RADC-TDR-64-287
Final Report

CLASSIFICATION SPACE ANALYSIS

TECHNICAL DOCUMENTARY REPORT NO. RADC-TDR-64-287

October 1964

Information Processing Branch
Rome Air Development Center
Research and Technology Division
Air Force Systems Command
Griffiss Air Force Base, New York

Project No. 4594, Task No. 459401

(Prepared under Contract No. AF 30(602)-3342 by Peter G. Assorio, Department of Psychology, University of Colorado, Boulder, Colorado.)

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Qualified requesters may obtain copies from Defense Documentation Center.

Defense Documentation Center release to Office of Technical Services is authorized.

Do not return this copy. Retain or destroy.

FOREWORD

A significant portion of the analyses described in this report were performed at the University of Colorado Graduate School Computing Center. In most cases, the programs used in the analyses were developed at the Western Data Processing Center (UCLA) for use in other linguistic research now in progress.

The data analyzed for this report was collected by the Air Force under separate contract No. AF 30(602)-2992.

Keywords: Documentation, Indexing

ABSTRACT

A conceptual approach to linguistic data processing problems is sketched and empirical illustrations are presented of the major software components - indexing, storage, and retrieval - of a document processing system which offers, in principle, the advantages of complete automation, unlimited cross-indexing, effective sequential retrieval, subdocumentary indexing reflecting heterogeneity of subject matter within a document, and a procedure for automatically identifying retrieval requests which would be inadequately handled by the system.

The indexing scheme, designated as a "Classification Space" consists of a Euclidean model for mapping subject matter similarity within a given subject matter domain. A scheme of this kind is empirically derived for certain fields of Engineering and Chemistry. A set of five related empirical studies provide convincing evidence that when appropriate experimental procedures are followed a very stable C-Space for a given content domain can be constructed on a surprisingly small data base.

Other empirical studies demonstrate specific computational procedures for effective automatic indexing of documents in a C-Space, using a relatively small system vocabulary. One study demonstrates that a C-Space maps subject matter relevance as well as subject matter similarity, and thereby promotes effective sequential retrieval; this result is also shown under conditions of automatic indexing.

Negative results are found in an attempt to use the structural linguistic distinction of subject and object as a means of improving techniques for automatic indexing.

PUBLICATION REVIEW

This report has been reviewed and is approved. For further technical information on this project, contact

Approved: Robert N. Ruberti
ROBERT N. RUBERTI
Project Engineer
Info Processing Branch

Approved:
ROBERT J. QUINN, JR.
Colonel, USAF
Ch., Intel and Info Processing Div.

FOR THE COMMANDER:
IRVING J. GABELMAN
Chief, Advanced Studies Group

TABLE OF CONTENTS

1.0 Introduction

Every civilization develops ways of preserving and transmitting the intellectual products--theoretical, technical, artistic--which it has accumulated through invention or borrowing. For some generations, now, we have been familiar with the Library as the principal institution through which this was accomplished, primarily through its relation to other institutions (e.g., Science, Education) which are more directly concerned with the construction and utilization of such products. Ideally, the Library makes its contribution by providing central, permanent storage for permanent records and convenient, economical access to them when they are needed. It is commonly agreed that during the past generation the Library has become increasingly less effective in its contribution to education, science, and technology. Indeed, at the present time, the Library appears to constitute a most critical bottleneck in relation to these other areas.

1.1 The Library Problem

Within the Library, it is the subject matter indexing, controlling access to documents, which is the distinctive and fundamental mechanism for the preservation and transmission of records. Briefly, the traditional subject matter indexing involves the categorization (by specially trained persons) of documents in accordance with some decision(s) as to what each document "is about". Documents are assigned to specific subject matter categories. Correspondingly, the User who wants access to documents in the Library goes through the following steps:

(a) He generates a criterial description of the kind of documents wanted, e.g., "something that tells me how a strut is braced" or "all about cadmium batteries".

(b) Calling on his knowledge about the Library indexing categories, he translates his criterial description into some number of secondary descriptions the sole utility of which is that they employ the library indexing categories and consequently, do succeed in making a selection of some kind from the totality of documents in the library. For example, the User who wants to know how a strut is braced may confine his search to documents listed under "Aeronautics", "Aerodynamics," "Aircraft Structure", "Wing Theory", "Stability and Control", and "Aircraft Design".

(c) Having decided on one or more indexing categories, he looks at all the entries under each category and does so in the order in which they are listed, since there is no other ordering available to him which is not equally arbitrary. Sometimes other categories of search are suggested by an initial search (e.g., "see: History of Flight"). Documents which are not listed under the categories which are searched do not become available to the User.

(d) Upon examining each title listed under a chosen indexing category, together with whatever descriptive material accompanies the listing, he makes a decision as to the likelihood that a document having that title and description contains what he wants. If the document is heterogeneous in content (e.g., a journal) it may be necessary to gain access to the document in order to make a preliminary decision. If, in addition, the document is a lengthy one (e.g., a textbook), extensive inspection may be required.

(e) He obtains temporary possession of the documents he selects (if some-one else has not already done so). This being accomplished, it may still require considerable searching through his selection of documents in order to identify relevant passages and assemble the desired information.

Thus, from the User's point of view, the Library Problem has the following elements:

(a) The descriptions which operate in selecting documents are only incidentally related to his criterial description, so that he has no general assurance that if the information he wants is in the Library he will get it, and he can estimate only crudely, if at all, what proportion of the relevant material in the Library he has acquired.

(b) The necessity of examining all the titles under a given category is so burdensome that for a given level of desperation it keeps to a minimum the number of categories searched.

(c) The title of a document, even with the supplementary information usually found in index files, is likely to be an insufficient guide as to the content, especially when the content is heterogeneous.

(d) Given the amount of searching required in the documents, on top of the amount of search required for the documents, theuse of the Library is so costly in terms of time and effort that the User is drastically limited in the amount of information he can afford to acquire in relation to any given project.

For a number of reasons, the Library Problem has become particularly acute in the area of science and technology. Among the major factors contributing to the seriousness of the problem, we find the following:

(a) The massive and ever-increasing volume of documents being produced and requiring indexing, storage, and access processing.

(b) The consequent premium on precision and selectivity of retrieval in order to make the use of information an economically feasible proposition.

(c) The premium on the immediate availability of new information to the Users to whom it is of interest.

(d) The accelerated evolution of multiple overlap and interrelatedness among the scientific and technical fields the names of which form the major basis for subject-matter categories; it is this feature, more than any other, which makes the "correct" assignment of a document to a specific one, two, or N indexing categories increasingly problematical.

(e) The premium on completeness--that is, the general reluctance to accept as adequate any selection of documents which does not contain "all" the documents which are relevant to the User's search.

1.2 Efforts Toward a Solution

The advent of high-speed, large-storage computers has generated intensive efforts to reduce the problem of the Library. There are, at the present time, several techniques for automatic or partially automatic document storage and retrieval which have demonstrated at least a moderate degree of success. Understandably, considerable effort is currently being devoted to the further improvement of these techniques.

A technical solution to the Library problem would appear to involve at least the following achievements:

(a) Complete cross-indexing of all documents with respect to indexing categories

(b) A retrieval process which incorporates an effective ordering principle for determining the order in which documents are retrieved

(c) Completely automatic indexing and retrieval. "Completely automatic" signifies that in a functioning LDP system there are no human links; instead, the human contribution is limited to such boundary conditions as (1) seeing to it that the documents are in a form physically suitable for computer processing, (2) presenting retrieval requests, (3) the initiation of the system, and (4) maintainance and monitoring operations.

A fundamental solution to the Library problem would involve, in addition, the following:

(d) Retrieval selection in accordance with an objective principle which effectively simulates the User's criterial description. This carries some implications in regard to the total scope of the system.

By and large, except for some of the most recent innovations which tend toward some kind of semantic analysis, existing approaches to the problem are technical efforts directed primarily toward effective cross-indexing and retrieval order. Collectively, they represent an intensive examination and exploitation of the physical parameters of the records (documents) which are to be indexed, stored, and retrieved. Word shapes, word sequences, and frequencies of their occurrence, co-occurrence, and contingent occurrence provide examples of such parameters.

Yet it seems clear that (a) the essential features of the Library Problem involve the functional parameters of these records, and (b) these parameters can be translated into physical parameters only within a descriptive context which extends beyond these intellectual products and deals with their production and consumption as well.

In the present study an attempt has been made to take account of the sources and uses of scientific and technical messages in a conceptually non-trivial and practically significant way. The specific goals of the efforts described in Section 2 have been

(a) To identify aspects of science and technology which are relevant to the Library Problem;

(b) To provide the empirical groundwork for a specific LDP system which would contribute to the present state of the art by virtue of offering complete automation, complete cross-indexing, and effective sequential retrieval;

(c) To present methodological considerations and empirical evidence in regard to the general feasibility of such an approach; and

(d) To contribute toward a general solution to the Library Problem by illustrating the effective implementation of criterial descriptions in the very limited, but perhaps uniquely important, case where the criterial description comes close to being purely a subject-matter description.

2.0 Psycholinguistic Studies for Automatic Linguistic Data Processing*

Over a period of ten months a series of psycholinguistic studies was undertaken with the aim of providing the methodological and empirical basis for a functioning storage and retrieval system for selected areas of science and engineering. A total of five separate studies was involved. These studies are characterized briefly below and are presented in detail in the sections which follow.

(1) The Classification Space Study: This study generates the software for an indexing system for certain fields of Chemistry and Engineering. The indexing schema consists of a geometric model for the subject matter domain in question.

(2) The Stability Study: In this study, evidence is presented in regard to the adequacy of the empirical base for the Classification Space Study.

(3) The Vocabulary Study: Here, evidence is presented in regard to the system dictionary requirements which would be encountered in implementing the C-Space index.

(4) The Relevance Ranking Study: Here, evidence is presented in regard to the effectiveness (validity) of the simplest sequential retrieval option which is made possible by C-Space indexing.

(5) The Grammatical Study: This represents an attempt to improve C-Space indexing by making use of some gross structural features of the linguistic data.

* The collection of data for these studies was accomplished with the substantial assistance of Mr. Ronald Taylor, Research Engineer, and Dr. George Motherwell, Research Linguist

2.1 Pragmatic concepts for subject matter

In this section, attention is directed to some relevant aspects of science and subject matter and, by reference to these, certain concepts are defined as technical terms for later use.

1. The production and consumption of scientific messages are primarily socio-cultural phenomena rather than spatio-temporal, though they are the latter also.

2. These socio-cultural phenomena are institutionalized, the institutions being those socio-historical entities which are commonly identified as "fields of knowledge". What we refer to as "science" when we speak of storing and retrieving scientific documents is a more or less disparate collection of fields of knowledge. Expressions such as "Chemistry", "Electronics", and "Biosynthesis" may be used to identify either a field of knowledge or the corresponding subject matter.

3. Thus, any field of knowledge has two distinct types of constituent. The first is a set of statements and specific concepts (e.g., "dipole moment", "indeterminacy", etc.) which, collectively, form thecontent of the field at a given time. The second is a set of specific people performing those distinctive tasks which, collectively, form the socio-historical reality of that field at a given time.

4. As with any historically existing individual (e.g., a table, a mountain, a nation, an army), a field of knowledge cannot be defined into existence, nor can its characteristics be arbitrarily assigned. Rather, it must be identified empirically and characterized a posteriori. There is, for example, no field

of knowledge in which "transistor", "empathy", and "acetylcholine" are all significant technical terms. (But there might have been, and there may yet be, such a field.) Further, it is currently the case that no existing theory or empirical generalization is of significant practical value in predicting the emergence of new fields of knowledge or in making inferences to as-yet-unexamined attributes of existing fields of knowledge.

5. In general, there are two major sources of criteria for identifying fields of knowledge. Some fields are identified primarily on an academic basis and are coherent because they are associated with distinctive theories, methods, or concepts and sometimes, also, because they form traditional curricular subdivisions. Other fields are defined primarily on a professional basis and are coherent because there is a set of practitioners who identify themselves distinctively (Electrical Engineer, Clinical Psychologist, Actuarial Accountant), communicate to one another on this basis, sustain a distinctive professional literature, etc. There is a strong tendency for the more general or extensive fields to be identified primarily on an academic basis and for the most specific fields to be identified primarily on a professional basis. This is to be expected, since the range of activities corresponding to the more general fields, e.g., "Psychology", exceed the range of individual human competence.

6. Fields of knowledge do not have a general property of divisibility. It is true that there are some hierarchical inclusion relationships among some fields, but it is not generally the case that a part of the subject matter of a field, A, is the entire subject matter of some other field, B. The reason is that the range of content of a given field is determined by the range of activity and interest of a specific set of people, namely,

the practicing scientists or engineers in that field. Without such a body of practices there is no field at all, and subdividing the content of a field provides neither a guarantee nor a presumption that there is any corresponding body of practices or practitioners.

7. The basic relationship between statements and fields of knowledge is the pragmatic relation of relevance. (This is a logical statement, not an empirical assertion.) A statement is relevant to a field (a) if it is generated by a practitioner as a part of his professional activity, or (b) if practitioners in the field would accept the statement as having been generated in this way. (Here, "is relevant" is to be understood as "has a practically significant degree of relevance".) Concomitantly, we can characterize individual technical terms (or their associated concepts) as being more or less relevant to a given field accordingly as they are used or may be used in statements which are more or less relevant to the field. A reasonable paraphrase for "relevant" would be "useful toward some (one or more) of the (more or less important) goals which are implied by the practitioner's professional activities". However, it is by no means clear the the pragmatic concept of "relevance" could be translated into other terms which were not likewise pragmatic concepts and did not equally require explication (cf "useful," "important).

8. That a certain technical term or concept has a certain degree of relevance to a certain field of knowledge is a significant attribute of that term and of that field. For both, it is a contingent, time-dependent attribute.

9. Molar discursive units such as paragraphs, chapters, documents, etc., are more or less relevant to a given field accordingly as the statements and concepts contained in them are more or less relevant to the field.

10. Within the domain of science and technology, the information potential of a statement (or a technical term, or a molar discursive unit) is summarized uniquely by the spectrum of its degree of relevance to the different fields of knowledge. Conversely, the subject content of a field of knowledge is uniquely summarized by the spectrum of the degree of relevance which the different terms have to that field.

11. The subject matter similarity of two technical terms (or larger units) is defined as the degree of similarity of their information potentials. Conversely, the subject matter similarity of two fields of knowledge is defined as the degree of similarity of their subject contents. These attributes inherit the contingent and time-dependent aspects of the basic relevance relationship.

12. Criterial evidence for the degree of relevance of technical expressions, statements, or larger units to a given field of knowledge can be obtained only from the practitioners in that field. This statement is to be understood not as an unfortunate practical consequence of the limitations mentioned in paragraph 4, above, but rather, as a necessary consequence of the concept of "field" presented in paragraph 3.

2.2 The Classification Space Study

This was the major project, in terms of data collection, among the five studies described in this report. The aim of the study was to provide the basis for an LDP system which would be sufficiently extensive to have some practical value in an operational setting. The general rationale for the C-Space Study is the following:

a. Given a set of technical expressions and a set of fields of knowledge, numerical estimates of the degree of relevance of each term to each field can be obtained by polling informants who are competent in these fields.

b. Given acceptable sampling of terms, fields, and informants from the total content domain under investigation, the product-moment correlation between any two fields, computed on the basis of the relevance measures referred to above, provide acceptable estimates of the subject matter similarity of that pair of fields.

c. A correlation matrix representing the intercorrelations among M variables can be factor analyzed and the result is a k-dimensional Euclidean space in which is embedded the collective scope of the similarity relationahips among the M variables. In the k-space the variables are represented as a configuration of M vectors fanning out from the origin. This configuration has the property that the angle between any two vectors is directly proportional to the correlation between the two corresponding variables.

d. If the M variables are fields of knowledge and the correlations are good estimates of subject matter similarity, the K-space represents the scope of all the subject matter which is common to at least two of the M fields. Subject content which is uniquely associated with one field will not be represented in the k-space. If a field has a substantial proportion of unique content, so that it is poorly represented in the k-space, it can be accomodated, for most purposes, by adding a new reference axis which is orthogonal to the other k axes and is uniquely associated with that field. Thus, the total subject content of a set of M fields of which r are substantially unique can be represented in N-space ($N = k + r$).

e. The degree of relevance of a technical term (or larger discursive unit) to a field may be represented as the projection of a "term vector" on the field vector. If the degree of relevance of the term to each of the M fields is known, the projections of the term vector on each of the M field vectors is known, and since the orientation of each field vector with respect to the reference axes is known, the projection of the term vector on each of the reference axes can be estimated. The estimation of these projections is equivalent to assigning the term to a specific location in the N-space.

f. There is a one-to-one relation between the coordinate values assigned to a term in the N-space and the set of projections of the corresponding term vector on the M fields. The latter set represents the information potential of that term with respect to the content domain defined by the M fields. Thus, the informational potential of the term is uniquely represented by its location in the N-space, and the N-space can now be seen as a coherent descriptive system for classifying discursive units according to their information potential with respect to a particular content domain. It is for this reason that an N-space of this kind is referred to as a "Classification Space" and serves as a system for subject-matter indexing.

g. An important property of a C-space is that (omitting qualifications which may stem from experimental or measurement difficulties) the distance between two discursive units located in the space is directly proportional to the measured degree of subject matter similarity of the two units.

h. If one of two discursive units located in a C-space has the pragmatic status of a retrieval request and the other is a sentence or larger unit (i.e., is a candidate for retrieval) then, within a limited distance range over which Users discriminate degrees of relevance, the C-space distance from the second unit to the retrieval request is a monotonic function of its degree of relevance to that retrieval request if the latter represents a purely subject-matter criterial description on the part of the User making the request. (Relevant empirical evidence is provided by the Relevance Ranking Study.)

i. The foregoing provides an effective procedure for sequential retrieval of items indexed in a C-Space: items are retrieved in the order of their C-Space distance from the retrieval request. The simplest case corresponds to a spherical search volume in the C-Space. Other options may be adopted. For example, a retrieval request may be permitted to give greater weight to some coordinate axes in the C-Space and thus define ovoid, cylindrical, rectangular, or quite irregular search volumes.

j. The location of a discursive unit in a C-Space based on M fields of knowledge is a more powerful classificatory resource than would be provided by the complete cross-indexing of that unit with respect to the M fields considered as discrete indexing categories. This is because the C-Space is interpretable throughout its range and not merely in those regions which coincide with the field vectors. Such a result may appear paradoxical in view of the general non-divisibility of fields of knowledge. However, this discreteness merely reflects the coherence of the field as such and is not a property of the logical space into which a collection of such fields is mapped. The fact is that (a) new fields of knowledge do evolve over time and are related to pre-existing fields in ways not fundamentally different from the ways in which the earlier fields are inter-related and (b) the derivation of a given C-Space may involve the neglect of an existing field which, if it had been included, would occupy a currently "empty" region of the C-Space. These facts make it quite clear that the notion of a "possible field" is not a mere verbalism but is a legitimate part of the C-Space concept. Or, the C-Space can be interpreted directly as a mapping of information potentials, and in this map the regions occupied by the field vectors have no special significance because the field vectors as such have no special significance--their initial function has been taken over by the reference axes.

k. In summary, a geometric model for subject-matter indexing carries with it two of the three elements of a technical solution to the Library problem. The third--complete automation--will be discussed in connection with the Vocabulary Study. As might be anticipated, the power implied by the three elements is not to be gained without significant cost.

2.2.1 Procedures

a. Content domain: The content domain for the Classification Space Study was defined by the selection of four major fields for investigation. These were Electrical Engineering, Aeronautical Engineering, Physical Chemistry, and Biochemistry.

b. Identification of fields: Within each of the four major areas a survey was made by one or more prima facie competent persons (See Appendix C) in order to identify the fields of knowledge falling within these areas. In most cases, these experts were graduate students who had passed their qualifying examinations and could thus be expected to have a broad acquaintance with 'he literature and fields in their general area of specialization. In the case of Electrical Engineering, the work was performed by a group of research engineers. In each area an attempt was made to identify a small number of broad and jointly exhaustive fields and a large number of very specialized fields. This procedure was adopted in the interest of (1) maximum differentiation of the total content domain and (2) minimization of complications associated with "unique content" (of Section 2.2-d).

c. Selection of fields: Approximately 250 fields were generated by the foregoing procedure. Because of the limitation on the number of variables which can be accommodated by currently available computer programs for factor analyses, the number of fields selected for empirical study was set at 130. Reduction was accomplished by making a forced-choice apriori assignment of specialized fields to the broader fields and eliminating specialized fields in such a way as to preserve at least one specialized field for each broader field. The list of 130 fields is given in Table 1.

d. Selection of corpus: The experts who identified the fields were also assigned the task of selecting a corpus. The general instructions were to try to sample textbooks, journals, and government documents within the last five years for each field. Specific criteria were that each selection should consist of at least six consecutive paragraphs, the whole of which was clearly

Table 1

C-Space Fields

*	1	Electric Machinery
	2	Power Transmission
*	3	Instrumentation
*	4	Radar
*	5	Field Theory
*	6	Audio Engineering
*	7	Power Generation and Distribution (excluding electronic power systems)
*	8	Solid State Engineering
*	9	Telephony
*	10	Aircraft Structures
*	11	Aerodynamics
	12	Aircraft Design
	13	Air Properties
*	14	Beam Theory
*	15	Catalysis
*	16	Self-consistent Field Theory
*	17	Fluctuations and Brownian Movement
*	18	High Energy Nuclear Chemistry
*	19	Dipole Moment and Polarizability
*	20	Drugs and Poisons
*	21	Biosynthesis
*	22	Structural Polysaccharides
	23	Simple Lipids
*	24	Enzymes
*	25	Circuit Theory
*	26	Control Engineering
*	27	Electronic Data Processing
*	28	Communication Theory
*	29	Microwave Engineering
	30	Wire Communications
	31	Illumination Engineering
	32	Industrial Electronics
*	33	Radio Engineering
	34	Television Engineering
	35	Electrochemistry
	36	Electrophysics
*	37	Analogue Circuitry
*	38	Digital Circuitry
*	39	Computer Software
*	40	Microminiaturization (circuits)
*	41	Electronic Recording Systems
*	42	Non-linear Circuit Analysis
*	43	Linear Circuit Analysis
*	44	Feedback Control Systems
*	45	Decision Processes

Table 1 (continued)

*	46	Control Theory
	47	RF Techniques
	48	LF Techniques
	49	Transformers
	50	Motors and Generators
	51	Power Distribution
	52	Space Power Systems
	53	Transmission Lines
	54	Transducer Engineering
	55	Medical Electronics
	56	Radio Astronomy
*	57	Electromagnetic Fields
	58	Electric Fields
	59	Magnetic Fields
	60	Microwave Networks
	61	Telegraphy
	62	Telemetry
	63	Semiconductor Design
	64	Solid State Systems
*	65	Quantum Devices
*	66	Crystallography
*	67	Electron Tubes
*	68	Detection Theory
	69	Modulation Theory
	70	Conductors and Insulators
	71	Electro-optics
	72	Piezo-electric Theory
	73	Electroacoustics
*	74	Antennas
	75	Batteries and Fuel Cells
	76	Propulsion
*	77	Stability and Control
	78	Aircraft Performance
	79	Wind Tunnel Evaluation
	80	Fluid Statics
	81	Incompressible Flow
	82	Radiation in Atmosphere
	83	Foundations of Thermodynamics and Fluid Dynamics
	84	Transonic Flow
	85	Re-entry Methods
	86	Ablation
*	87	Properties of Materials
	88	Airfoil and Wing Theory
	89	Rockets
	90	Fuels
	91	Exotic Methods
	92	Dynamic Stability
*	93	Static Stability
	94	Propellers, Gears, and Control Mechanisms

Table 1 (continued)

	95	Maneuvering Flight
	96	Mach Number Effects and Reynolds Number Effects
*	97	Kinetics
*	98	Spectroscopy
*	99	Thermodynamics
*	100	Quantum Chemistry
*	101	Statistical Mechanics
*	102	Nuclear and Radiochemistry
	103	Electrochemistry and Magnetochemistry
	104	Exotic Fuels
	105	Liquid Kinetics
*	106	Transition State Theory
*	107	Photochemical Reactions
	108	Surface and Colloid Chemistry
*	109	Chemical or Phase Equilibrium
*	110	Valence Bond Theory
	111	Theory of Dense Fluids
	112	Radiation Chemistry
	113	Magnetic Susceptibility
*	114	NMR and EPR
	115	Optical Pumping
	116	Rotational (microwave) Spectroscopy
	117	Structural Biochemistry
	118	Biochemical Energetics
	119	Biochemical Kinetics
	120	Biochemistry of Diseases and Anomalies
	121	Biochemistry of Nutrition
*	122	Biochemical Genetics
	123	Experimental Biochemistry
*	124	Biochemistry of Metabolism
*	125	Sugars
	126	Nucleoproteins and Nucleic Acids
*	127	Amino Acids and Structural Proteins
	128	Coenzymes and Activators
	129	Vitamins and Hormones
	130	Pigments

part of the literature of the field in question. Six such references were obtained for each field, and each paragraph and sentence was identified by number.

e. Identification of technical expressions: For each of six numbered paragraphs in each of six references for each field, the experts were asked to underline all the technical expressions in these passages. Either words or phrases were acceptable as technical terms. Subsequently, the corpus was submitted to non-experts whose task was to indicate all the non-underlined words which they did not regard as part of ordinary English. This procedure revealed the fact that it was not at all uncommon for technical terms from field A to appear in literature clearly belonging to field B and in a majority of cases (e.g., numbers, special symbols, names, and equations) arbitrary decisions to underline or not were made on an intuitive basis. For example 1000 KW was included (power generation) and 31° was not (aerodynamics); "Dalziel criterion" and "Laplace" were included, but most names were not; "E → O" probably should have been included, but was not.

f. Selection of technical expressions: An essentially random selection of two technical terms from each reference was made. (The selection table which was genuinely random for the first 24 fields was reapplied forward and backward to successive blocks of 24 fields). In addition, one sentence was selected at random from the corpus for each field. Thus the list of "terms" used in the C-Space study consisted of twelve technical expressions and one sentence from each of the 130 fields.

g. Selection of informants: The C-Space Study was originally designed to make use of a minimum of three expert informants from each of the (130) fields selected for study, with equal representation of academic faculty, graduate students, and other professional personnel. However, considerable difficulty was encountered in obtaining the services of some classes of expert informants, and the result was that graduate students from UCLA, Stanford University, and the University of Colorado made up approximately 75% of the informants and the remaining 25% consisted of other professional personnel. Moreover, the minimum number of three informants per field was achieved for only 60 of the 130 fields. These 60 are indicated by asterisks in Table 1.

h. Apparatus: The C-Space terms were presented to informants in the form of a 141-page booklet containing twelve terms or six sentences per page, each item in the format shown in Appendix A. Over the total sample of informants, the order of presentation of the material was approximately counterbalanced in blocks of 24 consecutive pages in the "normal" 141-page sequence.

i. Instructions: The written instructions to the informants are,given in Appendix A. However, experience has shown that written instructions alone are frequently not effective. The overall introduction of the informants to the task was normally accomplished in an orientation session for groups of informants ranging from two to fifteen in number. This session usually lasted from half an hour to one hour and included (a) presentation of the written instructions, (b) preliminary practice ratings (c) verbal amplification of the written instructions (d) question and answer periods in regard to the task, and (e) a brief explanation of the nature and purpose of the C-Space Study. In general, each informant was instructed to make a direct judgment as to the degree of relevance of each term to his field of competence.

2. 2. 2. Results

Sixty C-Space fields were inter-correlated on the basis of mean estimates (i. e. averaging the judgments of the informants for a given field) of the relevance of 1548 technical expressions and 130 sentences to those fields. From the correlation matrix 26 factors were extracted by means of Comrey's Minimum Residual method of factoring. When these were rotated to a Varimax criterion thirteen of the factors retained appreciable loadings; the remaining thirteen factors may be regarded as uninterpretable residuals. The thirteen interpreted factors summarized in Table 2 account for 70 per cent of the total variance and 94 per cent of the common variance of the 60 variables analyzed. The summarization of the factor matrix is achieved by listing the fields separately for each factor in the order of decreasing magnitude of their loadings (i. e. projections of the field vectors on the reference axes) on that factor and omitting those fields which have loadings of less than .400 and therefore do not contribute appreciably to the characterization of the factor.

Inspection of the summarized factor results shows that the configuration of field vectors in the C-Space is eminently in accord with expectations based on a general understanding of the nature of the individual fields. There is no instance of two fields which are prima facie quite different in content being represented in the C-Space as having a high degree of similarity. This fact provides one measure of assurance as to the methodological adequacy of the

C-Space, since in the application of factor analytic techniques a common indication of conceptual or experimental inadequacy is the finding of paradoxical similarity relationships. It is also apparent, however, that many of the more specific differences among the fields have not been articulated within the C-Space, which reflects only common variance. Instead, these differences appear to be represented primarily by the pattern of specific variances for the fields. Of the 60 fields, more than one third have "unique" content which accounts for thirty to fifty per cent of their variance. Thus, the thirteen factors resulting from the present analysis represent only the common variance K-Space; a functional C-Space based on this analysis would require an N-Space of 30-35 dimensions.

It seems likely that the loss of some fine differentiation in this analysis is primarily due to sampling bias resulting in part from the controlled reduction of fields from 250 to 130, but more importantly from the uncontrolled reduction of fields from 130 to 60. Of the latter, 30 are Electrical Engineering fields and 16 are Physical Chemistry fields; only 14 are Aeronautical Engineering or Biochemistry fields. It is not surprising, therefore, that the EE and PC fields are distributed among ten dimensions whereas the BC and AE fields are encompasses in only three dimensions.

Table 2

Classification Space

Factor I Electronics

.866	Radio Engineering
.857	Analogue Circuitry
.829	Circuit Theory
.810	Communication Theory
.785	Microwave Engineering
.739	Radar
.738	Antennas
.735	Linear Circuit Analysis
.733	Non Linear Circuit Analysis
.728	Digital Circuitry
.703	Radio Frequency Techniques
.659	Electron Tubes
.653	Detection Theory
.639	Control Engineering
.629	Telephony
.601	Instrumentation
.590	Control Theory
.575	Electronic Recording Systems
.505	Feedback Control Systems
.494	Microminiaturization (circuits)
.484	Electric Machinery
.465	Solid State Engineering
.404	Electronic Data Processing

Factor II Subatomic Chemistry

.881	Quantum Chemistry
.796	Valence Bond Theory
.785	Self-Consistent Field Theory
.740	Spectroscopy
.654	Quantum Devices
.653	NMR and EPR
.624	Photochemical Reactions
.593	Dipole Moment and Polarizability
.495	Crystallography
.487	Nuclear and Radiochemistry
.423	High Energy Nuclear Chemistry

Table 2 (continued)

Factor III Biochemistry

.927	Enzymes
.897	Biosynthesis
.889	Biochemistry of Metabolism
.802	Sugars
.799	Biochemical Genetics
.785	Drugs and Poisons
.787	Amino Acids and Structural Proteins
.730	Structural Polysaccharides
.664	Catalysis

Factor IV Aircraft Structure

.897	Aircraft Structures
.867	Static Stability
.866	Beam Theory
.644	Properties of Materials
.401	Aerodynamics

Factor V Computer Software

.856	Computer Software
.762	Electronic Data Processing
.725	Decision Processes
.342	Control Engineering
.346	Detection Theory

Factor VI Molecular Dynamics

.784	Thermodynamics
.742	Chemical or Phase Equilibrium
.696	Statistical Mechanics
.627	Kinetics
.609	Fluctuations and Brownian Movement
.573	Transition State Theory

Factor VII Controls

.552	Control Theory
.548	Stability and Control
.534	Feedback Control Systems
.516	Control Engineering

Table 2 (continued)

Factor VIII Field Theory

.558	Field Theory
.444	Electromagnetic Fields
.336	Antennas
.305	Microwave Engineering

Factor IX Electric Machinery

.699	Electric Machinery
.657	Power Generation and Distribution

Factor X Aerodynamics

.538	Aerodynamics
.304	Stability and Control

Factor XI Nuclear and Radiochemistry

.694	Nuclear and Radiochemistry
.611	High Energy Nuclear Chemistry

Factor XII Solid State Engineering

.599	Solid State Engineering
.537	Microminiaturization (circuits)

Factor XIII Magnetic Fields

.364	Magnetic Fields
.353	Dipole Moment and Polarizability

2.3 The Stability Study

The concept of the subject content of a field of knowledge was defined with respect to a universe of technical terms. The terms used in the C-Space Study represent approximately ten per cent of the terms available in the 36-paragraph corpora associated with the set of 130 fields. (A spot check on the references for twenty Electrical Engineering fields revealed approximately 2300 technical terms, yet of this number only 240 terms and 20 sentences were used in the C-Space Study.) The terms used represent perhaps one per cent of the terms which might have been identified in a serious attempt to exhaust the "clearly relevant" literature.

Thus, it is of critical importance to obtain some empirical evidence as to the likelihood that a different literature or a different selection of terms from the corpus actually used would have given substantially different results.

Likewise, it was stated that for the C-Space a minimum of three informants per fields was required, but it would be quite appropriate to ask, "Was that enough? Would three or six or ten other informants have given the same results?"

The Stability Study consists of a set of five experiments designed to provide an empirical basis for answering questions as to the stability of factor results:

(a) When only the number of terms per field is varied

(b) When informants are different and a different selection of terms is made from the same corpus as (a).

(c) When informants and number of terms remain as in (b) but terms are from a different corpus.

(d) When judgments by informants in a given field are not averaged but are treated as separate measures.

(e) When the total sample of terms is qualitatively appropriate for some of the fields in the analysis and not for the others.

These relationships are shown in greater detail in Table 3. Except for the column headed "Analyses", the letter entries in Table 3 have no external reference; they merely indicate in which respects the conditions for the different experiments were alike or different. The "analysis" designations are those used in the descriptions of the experiments in the sections which follow.

2.3.1 Stability Experiment I

This is a study of the effect of varying the number of terms from each field within a C-Space factor analytic paradigm. Twenty-four fields from the 130 listed in Table 1 were selected for study. These twenty-four, shown in Table 4, represent random selections within the four major content areas, with the total number from each area fixed in advance. Nine fields were drawn from the Electrical Engineering fields and five each from the other three areas. Each field was represented by three informants except for fields 15, 18, and 24 (4 informants) and field 11 (five informants), so that a total of 77 informants were used. Of these approximately 80 per cent were graduate students, mainly from UCLA, and the remaining 20 per cent were research engineers or civil engineers in industry. The instructions to the informants were the same as for the C-Space Study, and likewise the physical format in which the judgments were obtained.

Table 3

Stability Experiments

			Term	Reference	Method of
Experiment	Analysis	Informants	Set	Set	Factoring
I	A	P	A_1	X	J
I	B	P	A_2	X	J
I	C	P	A_3	X	J
I	D	P	A_4	X	J
I	E	P	A_5	X	J
I	F	P	A_6	X	J
I	G	P	A	X	J
II	SE-II	P'	A	X	J
III	K	Q	B	X	H
IV	L	Q	C	Y	H
V	M	P	A'	X	H
V	N	P	A	X	H

Twelve terms were selected at random from each field corpus, making a total of 288 terms. For each of the 24 fields, the terms were identified by number according to the order of their selection from the field corpus. The numbers served only to identify the separate terms--no use was made of the order. Thus, "term 1 from each field" identifies a specific set of 24 of the 288 terms for which data was available. Seven different selections of terms from this pool were made. In each case, the twenty-four fields were inter-correlated on the basis of averaged ratings on the selected terms and the correlation matrix was factor analyzed by the same method as in the C-Space Study. The selections of terms are shown in Table 5.

The selections shown in Table 5 were designed to provide the least possible overlap of terms from one analysis to another. Thus, analysis A shares one term with each of B, C, and D; Analyses E and F each share two terms with B, C, and D; E and F are mutually exclusive, as are B, C, and D.

The results of the seven analyses are summarized in Table 6. All of the analyses show six major factors and one or more minor factors. In order to facilitate comparison the results of the separate analyses are not presented separately. Instead, the factors which most nearly resemble each other across the seven analyses are grouped together. In general, the summary retains only the field-factor combinations in which the loading of the field on the factor is greater than .400. Wherever a loading of less than .400 is shown, all higher loadings on that factor are shown. For each factor, the fields are listed in the order of decreasing loadings in analysis G.

On the whole the results demonstrate a remarkable degree of correspondence across all seven analyses. There is no appreciable trend in the

Table 4

Stability Experiment I Fields

* 1. Electric Machinery
2. Power Transmission
* 3. Instrumentation
4. Radar
* 5. Field Theory
6. Audio Engineering
* 7. Power Generator and Distribution
8. Solid State Engineering
9. Telephony
* 10. Aircraft Structures
* 11. Aerodynamics
12. Aircraft Design
13. Air Properties
14. Beam Theory
15. Catalysis
* 16. Self-Consistent Field Theory (SCF)
17. Fluctuations and Brownian Movement
* 18. High Energy Nuclear Chemistry (HENC)
19. Dipole Moment and Polarizability
20. Drugs and Poisons
* 21. Biosynthesis
22. Structural Polysaccharides
23. Simple Lipids
* 24. Enzymes

Table 5

Term Selections for S.E. I

Analysis	Term Selection					
A	1	2	3			
B	1	4	7	10		
C	2	5	8	11		
D	3	6	9	12		
E	1	3	5	7	9	11
F	2	4	6	8	10	12
G	All 12 terms from each field					

Table 6

Stability Experiment I Results

Factor I Atomic and Subatomic Dynamics

Analysis							Field
A	B	C	D	E	F	G	
.672	.720	.717	.810	.804	.790	.796	Dipole Moment and Polarizability
.450	.712	.730	.543	.746	.710	.722	Self-Consistent Field Theory
.417	.433	.612	.539	.583	.600	.610	Solid State Engineering
.541	.533	.335	.652	.518	.627	.599	High Energy Nuclear Chemistry
	.552	.456	.426	.642	.508	.569	Field Theory
.415	.537	.352	.769	.574	.508	.541	Fluctuations and Brownian Movement
.467	.281	.223	.475	.382	.384	.413	Catalysis
	.300	.327		.377	.332	.354	Radar
		.297					Instrumentation

Factor II Electronic Machinery

Analysis							Field
A	B	C	D	E	F	G	
.830	.749	.769	.754	.703	.782	.754	Audio Engineering
.802	.708	.723	.522	.630	.714	.675	Instrumentation
.577	.637	.743	.569	.651	.656	.643	Telephony
.735	.638	.677	.505	.577	.677	.601	Radar
.689	.550	.480	.454	.445	.501	.483	Solid State Engineering
.522	.520	.441		.359	.446	.377	Field Theory
.606							SCF

Factor III Molecular (Fluid) Dynamics

Analysis							Field
A	B	C	D	E	F	G	
.848	.871	.800	.863	.862	.890	.884	Aerodynamics
.844	.821	.787	.829	.817	.812	.882	Air Properties
.742	.784	.598	.800	.725	.764	.740	Aircraft Design
.635	.606	.564	.420	.487	.587	.536	Fluctuations and Brownian Movement

Factor IV Aircraft Structure

Analysis							Field
A	B	C	D	E	F	G	
.881	.895	.921	.885	.888	.896	.898	Aircraft Structures
.873	.878	.832	.872	.864	.884	.879	Beam Theory
.456	.444	.652	.409	.516	.429	.481	Aircraft Design
		.461					Aerodynamics

Table 6 (continued)

Factor V Biological Chemistry

Analysis							Field
A	B	C	D	E	F	G	
.921	.911	.876	.923	.912	.904	.914	Biosynthesis
.907	.913	.867	.887	.904	.899	.903	Enzymes
.782	.813	.799	.815	.760	.841	.815	Drugs and Poisons
.612	.637	.718	.709	.649	.708	.694	Simple Lipids
.623	.645	.264	.688	.527	.584	.557	Catalysis
.502	.500	.495	.578	.458	.572	.515	Structural Polysaccharides

Factor VI Electric Machinery

Analysis							Field
A	B	C	D	E	F	G	
.806	.818	.833	.877	.859	.810	.830	Electric Machinery
.800	.771	.831	.838	.857	.784	.824	Power Transmission
.782	.806	.808	.852	.821	.807	.807	Power Generation and Distribution
.428	.372	.354	.537	.432	.398	.425	Telephony
	.344	.249	.404	.342	.311	.323	Instrumentation
		.351	.351	.365			Audio Engineering

Factor VII Field Theory

Analysis							Field
A	B	C	D	E	F	G	
.520		.447	.705	.352	.370	.372	Field Theory
		.388	.486	.359	.189	.331	Radar
			.603				Self-Consistent Field Theory

Factor VIII Minor Factor

Analysis							Field
A	B	C	D	E	F	G	
	.325			.406			Simple Lipids
	.253						Drugs and Poisons
		.587					Catalysis
		.548					High Energy Nuclear Chemistry

results--the three-term analysis does not approximate the 12-term analysis any less closely than the 4-term or 6-term analyses do. It is also the case that, as in the C-Space results, the grouping of fields which emerges from the analysis is highly satisfactory from the point of view of a general understanding of the nature of these 24 fields. In fact, there is a decided resemblance between the present results and the C-Space results. To the seven factors identified in Table 6 there correspond the C-Space factors 2, 1, 10, 4, 3, (9 and 6) and 8. In the light of the present evidence, therefore, it would seem that the possibility that the C-Space structure is substantially biased by virtue simply of a too-small data base is not a significant issue.

On the positive side the fact that the ratio of invariance (of results) to size (of corpus examined) which is demonstrated here is exceptionally good provides some evidence that the pragmatic conceptualization outlined above does enable us to tap significant functional attributes of linguistic data.

2.3.2 Stability Experiment II

This experiment was designed to provide data relevant to the C-Space Study requirement of a minimum of three informants per field. The data for the present analysis was the data used for analysis G in SE-I, i.e. twelve terms per field for each of the 24 fields. In the present case, each of the 77 informants was treated as a separate variable, so that the analysis was an analysis of 77 informants rather than of 24 fields. As in all the previous analyses, the Minimum Residual method of extraction and the Varimax criterion for rotation were used. The factor results are summarized in Table 7, with each informant separately identified (e.g., 1-Enzymes, 2-Enzymes, etc.).

An examination of Table 7 suggests that six of the sixteen factors can be coordinated roughly with six of the seven factors consistently shown in the SE-I analyses. These are, respectively, factors I, II, V, III, IV, and VII. However, without the prior availability of the SE-I results, these factors could not have been so clearly interpreted. For example, the finding of 2-Drugs and Poisons and 1-Drugs and Poisons together with 3-Dipole Moment and 3-Self Consistent Field Theory at the top of the list of informants associated with Factor I appears quite paradoxical until we recall that (a) in SE-I, Dipole Moment and SCF were associated primarily with Factor I Atomic and Subatomic Dynamics and (b) the fine chemical structure of drugs and poisons is generally a critical factor in their biological activity. Given this much, we can then assimilate 1-Telephony to the interpretation of the present Factor I as Atomic and Subatomic Dynamics.

An analogous "decoding" of the grouping of Beam Theory, Catalysis, and Aerodynamics on Factor V gives less satisfying results, but we are reminded that (a) in aircraft work, Beam Theory deals primarily with metal beams and (b) catalysis is an important aspect of metallurgy. Even more resistant to interpretation is the grouping of 1-Aircraft Structures, 3-Audio Engineering, 2-Power Generation and Distribution, and 1-Radar on Factor IX. Further, although the finding of 3-Henc and 4-Henc as the sole representatives of Factor XIII may be considered encouraging, it becomes impossible to interpret this factor with any degree of confidence when we find, in addition, 1-Henc and 2-Henc together with 3-Fluctuations as the sole representatives of Factor XVIII. Again, factors IV, VIII, and IX all appear to be heterogeneous "Electronics" factors of some kind, but it is difficult to say what kind or to characterize the differences among them.

Thus, on the whole, the present results reflect a marked lack of agreement among informants within fields, and, in contrast to the SE-I results, they do not present a well-structured configuration which makes sense in relation to a general understanding of the nature of **the fields.** It would appear, therefore, that the requirement of three informants per field is not an excess of caution and, in light of results such as those for Henc described above, it seems likely that three informants does represent a minimum figure.

One explanation for the present findings is that each of the informants represents a very imperfect measuring instrument and that averaging scores of informants within fields is effective because the errors of measurement associated with each informant are independent of the errors of measurement associated with other informants and consequently tend to cancel one another in the averaging process. Such an explanation, however, will not account for the magnitude of the effect, observed. Inspection of the correlation matrix shows the average correlation between two informants from the same field to be somewhat less than .50. Under the "errors of measurement" explanation, this would imply that from 75% to 100% of the variance associated with the ratings of each informant was error variance. To expect errors of this magnitude to cancel out by averaging <u>three</u> measures would be to expect silk purses from sows' ears. Using some crude, but conservative inequalities, one may calculate that under these conditions it would require nine judges to reduce the error variance in the averaged ratings to the point where it was merely <u>equal</u> to the true variance. Moreover, on any account involving error-cancelling one would expect to find a sample-size effect, since the amount of

Table 7

Analysis by Individuals

Factor I Atomic and Subatomic Dynamics

.643	2 - Drugs and Poisons
.630	3 - Dipole Moment and Polarizability
.599	1 - Drugs and Poisons
.583	3 - Self-Consistent Field Theory
.578	3 - Fluctuations and Brownian Movement
.509	1 - Telephony
.491	3 - High Energy Nuclear Chemistry
.454	1 - Self-Consistent Field Theory
.420	2 - High Energy Nuclear Chemistry
.407	2 - Solid State Engineering

Factor II Molecular (Fluid) Dynamics

.864	3 - Aerodynamics
.842	4 - Aerodynamics
.827	5 - Aerodynamics
.825	2 - Air Properties
.783	2 - Aircraft Design
.774	3 - Aircraft Design
.753	3 - Air Properties
.691	1 - Aircraft Design
.677	1 - Air Properties
.658	1 - Beam Theory
.649	2 - Beam Theory
.510	2 - Aerodynamics
.489	1 - Aerodynamics
.482	1 - High Energy Nuclear Chemistry

Factor III Biological Chemistry

.855	3 - Biosynthesis
.817	3 - Enzymes
.814	4 - Enzymes
.795	2 - Biosynthesis
.775	1 - Enzymes
.756	3 - Drugs and Poisons
.727	1 - Biosynthesis
.726	1 - Structural Polysaccharides
.720	2 - Structural Polysaccharides
.685	3 - Simple Lipids
.584	1 - Simple Lipids
.574	2 - Simple Lipids
.565	2 - Enzymes
.535	3 - Catalysis
.411	3 - Structural Polysaccharides

Table 7 (continued)

Factor IV Electric Machinery

.861	1 - Solid State Engineering
.791	3 - Power Transmission
.775	2 - Power Transmission
.764	1 - Power Transmission
.725	1 - Instrumentation
.699	3 - Electric Machinery
.675	2 - Instrumentation
.661	2 - Solid State Engineering
.608	3 - Power Generation and Distribution
.471	3 - Telephony
.460	2 - Power Generation and Distribution
.455	3 - Instrumentation
.428	3 - Audio Engineering

Factor V Aircraft Structure

.820	3 - Beam Theory
.758	1 - Catalysis
.651	3 - Aircraft Structures
.567	1 - Aerodynamics
.544	2 - Catalysis
.541	2 - Aerodynamics

Factor VI

.495	2 - Enzymes
.420	2 - Structural Polysaccharides

Factor VII Field Theory

.710	3 - Radar
.688	1 - Field Theory
.479	2 - Field Theory
.475	2 - Radar
.449	1 - Audio Engineering
.355	2 - Dipole Moment and Polarizability

Factor VIII

.529	3 - Telephony
.418	1 - Power Generation and Distribution
.381	2 - Solid State
.368	2 - Audio Engineering

Table 7 (continued)

Factor IX

.608	1 - Aircraft Structures
.534	3 - Audio Engineering
.505	2 - Power Generation and Distribution
.475	1 - Radar
.416	3 - Solid State Engineering
.381	3 - Instrumentation

Factor X

.572	3 - Field Theory

Factor XI

.634	1 - Self-Consistent Field Theory
.483	2 - Self-Consistent Field Theory
.420	4 - Catalysis
.398	3 - Catalysis

Factor XII

.651	1 - Electric Machinery
.597	2 - Electric Machinery

Factor XIII

.638	4 - High Energy Nuclear Chemistry
.535	3 - High Energy Nuclear Chemistry

Factor XIV

.557	3 - Structural Polysaccharides
.382	1 - Simple Lipids

Factor XV

.429	1 - Fluctuations and Brownian Movement

Factor XVI

.472	1 - Beam Theory
.471	2 - Beam Theory

Table 7 (continued)

Factor XVII

.570	2 - Fluctuations and Brownian Movement

Factor XVIII

.357	2 - High Energy Nuclear Chemistry
.348	1 - High Energy Nuclear Chemistry
.253	3 - Fluctuations and Brownian Movement

actual cancelling would approach its expected value for increasing N.

A more parsimonious explanation follows directly from the pragmatic conceptualization presented in Section 2.1. In that section, fields of knowledge were described as socio-historical realities and practitioners were described as constituents of such fields. On this view it follows that we can attribute to an individual practitioner no more than a particular view of his field. It follows further that we may expect practitioners to disagree in their judgments and that an adequate characterization of the field is best approximated by the sum of their judgments rather than the lowest common denominator--disagreement is by no means equivalent to error variance. The "decoding" of factors I and V in the present experiment is consistent with this explanation.

2.3.3. Stability Experiment III

This is a study of the changes brought about in the factor structure shown in SE-I when two parameters of the experimental situation--the informants and the specific set of technical terms--were altered. For practical reasons, the new set of terms was restricted to twelve terms drawn from each of the ten fields listed in Table 8. Because of the limited availability of informants and the time pressures operating in the collection of data, it proved impossible, in SE-III and SE-IV, to meet the requirement of three informants for each field. For two of the fields, Telephony and Power Transmission, no informants could be obtained. Because of the complications resulting from these limitations on the extent and quality of the data, the presentation and discussion of the analysis and results of SE-III will be postponed until Section 2.3.5.

Table 8

Field Sources of Terms for SE-III, -IV, and -V

1. Electric Machinery
3. Instrumentation
5. Field Theory
7. Power Generation and Distribution
10. Aircraft Structures
11. Aerodynamics
16. Self-Consistent Field Theory
18. High Energy Nuclear Chemistry
21. Biosynthesis
24. Enzymes

Table 9

Number of Informants for SE-III and SE-IV

Number	Field
4	Electric Machinery
0	Power Transmission
5	Instrumentation
5	Radar
2	Field Theory
2	Audio Engineering
1	Power Generation and Distribution
1	Solid State Engineering
0	Telephony
7	Aircraft Structures
4	Aerodynamics
5	Aircraft Design
3	Air Properties
5	Beam Theory
4	Catalysis
5	Self-Consistent Field Theory
3	Fluctuations and Brownian Movement
4	High Energy Nuclear Chemistry
3	Dipole Moment and Polarizability
4	Drugs and Poisons
5	Biosynthesis
3	Structural Polysaccharides
1	Simple Lipids
5	Enzymes

2.3.4 Stability Experiment IV

The previous Stability Experiments were based on terms selected from the six six-paragraph references constituting the field corpus for each field, as described in connection with the C-Space Study. In the present experiment a new 36-paragraph corpus was selected for each of the ten fields listed in Table 8, making use of the same principles of selection as the preceding corpora. Technical terms in the corpus were identified in the same way as before. and twelve terms were selected at random from each of the ten fields listed in Table 8. This set of 120 terms was rated by the same informants who provided the ratings for SE-III. Approximately two-thirds of the informants were those who provided the C-Space ratings for the fields which were studied in the present experiment. The remaining third consisted of graduate students from the California Institute of Technology. The analysis of this data and a discussion of the results will be presented in the following section.

2.3.5 Stability Experiment V

The data collected for SE-III and SE-IV falls significantly short of the ideal in at least two major respects. First, the three-informant minimum was not attained for five of the 22 fields analyzed. Second, the terms which were rated with respect to the 22 fields were drawn from only ten of those fields. The first of these shortcomings would be expected to produce results similar to those of the analysis by individuals, i.e. the results of SE-II as contrasted with the results of SE-I. The second would be expected to introduce appreciable error in the orientation of field vectors with respect to one another and to both increase and decrease the apparent similarity among fields, especially the fields not represented by any terms in the set of terms rated by informants. Thus, between the tendencies to (a) increase random error, (b) generate spurious similarities, and (c) generate spuriously idiosyncratic relationships among fields, the data limitations in SE-III and SE-IV presented the possibility of serious distortions in the factor results for these two stability studies.

For this reason, an intermediate set of data was constructed. From the data used in SE-I analysis G (12 terms per field) a selection was made of the ratings of those terms which had been drawn from the ten fields shown in Table 8. Data for the fields of Telephony and Power Transmission was eliminated. Thus, the SE-V data matched the SE-III and SE-IV data in regard to fields analyzed and fields from which the terms were taken; it was not matched in regard to the number of informants in specific fields.

The 22 fields represented by the SE-V data were intercorrelated and the correlation matrix was factor analyzed, using the Maximum Likelihood method of extraction and the Varimax criterion for analytic rotation. This analysis is identified as Analysis M, and the results are summarized in Table 10.

Similar analyses were made of these 22 fields, using the SE-III data and the SE-IV data. These are identified as Analysis K and Analysis L, respectively, and the results are summarized in Table 10.

A similar analysis was made of the entire data (24 fields, 288 terms) used for analysis G in SE-I. This is identified as Analysis N, and the results are summarized in Table 10. Analyses G and N differ only in that the Minimum Residual method of factor extraction was used in the former and the Maximum

Table 10

Analyses K, L, M, and N

Factor I Atomic and Subatomic Dynamics

	Analysis K	L	M	N	K	L	Field
**	.879	.470	.808	.836			Self-Consistent Field Theory
			.705	.798	.721	.669	Dipole Moment and Polarizability
o			.400	.635	.639	.564	Solid State Engineering
**o	.418	.442	.548	.630			Field Theory
		.642	.434	.513			Fluctuations and Brownian Movement
*			.392	.545			High Energy Nuclear Chemistry
		.399			.515	.388	Air Properties
*						.395	Instrumentation

Factor II Electronic Machinery

	Analysis K	L	M	N	Field
o	.672	.832	.789	.777	Audio Engineering
**	.844	.763	.778	.544	Instrumentation
	.548	.527	.750	.431	Radar
oo				.665	Telephony
o			.664	.409	Solid State Engineering
* o			.475		Field Theory

Factor III Molecular (Fluid) Dynamics

	Analysis K	L	M	N	K	Field
**	.836	.734	.895	.893		Aerodynamics
	.252	.661	.873	.844	.576	Air Properties
	.521	.278	.727	.764		Aircraft Design
			.662	.556	.843	Fluctuations and Brownian Movement

Factor IV Aircraft Structure

	Analysis K	L	M	N	Field
**	.925	.951	.923	.912	Aircraft Structures
	.932	.862	.921	.915	Beam Theory
	.664	.838	.483	.465	Aircraft Design
		.535		.275	Aerodynamics

Table 10 (continued)

Factor V Biochemistry

Analysis					Field
	K	L	M	N	
**	.886	.854	.852	.923	Biosynthesis
**	.951	.928	.437	.909	Enzymes
	.855	.842	.676	.827	Drugs and Poisons
o	.680	.434	.457	.734	Simple Lipids
	.899	.739	.631	.548	Catalysis
	.775	.821	.680	.540	Structural Polysaccharides
	.471				Radar

Factor VI Electric Machinery

Analysis					Field
	K	L	M	N	
**	.808	.773	.828	.866	Electric Machinery
**	.840	.953	.833	.853	Power Generation and Distribution
oo				.843	Power Transmission
oo				.423	Telephony

Factor VII Radar

Analysis					Field
	K	L	M	N	
			.448	.710	Radar
* o			.331	.334	Field Theory

Factor VIII

Analysis					Field
	K	L	M	N	
			.464	.539	Catalysis
**			.583	.362	High Energy Nuclear Chemistry

Factor IX

Analysis					Field
	K	L	M	N	
			.391	.400	Fluctuations and Brownian Movement

Table 10 (continued)

Factor X

Analysis					Field
	K	L	M	N	
o	.625	.851	.718		Simple Lipids
			.385		Biosynthesis
			.365		Drugs and Poisons

Factor XI

Analysis				Field
K	L	M	N	
.468	.427			Structural Polysaccharides

Table 11

Communalities for Analyses K, L, M, and N

Analysis						Field
		K	L	M	N	
*		.739	.704	.906	.918	Electric Machinery
					.876	Power Transmission
*		.873	.850	.869	.782	Instrumentation
		.690	.605	.854	.988	Radar
*	o	.245	.277	.847	.691	Field Theory
	o	.515	.727	.767	.858	Audio Engineering
*	o	.797	.945	.901	.886	Power Generation and Distribution
	o	.594	.488	.784	.725	Solid State Engineering
					.843	Telephony
*		.963	.965	.940	.925	Aircraft Structures
*		.950	.901	.934	.924	Aerodynamics
		.885	.889	.932	.933	Aircraft Design
		.720	.765	.811	.774	Air Properties
		.887	.870	.911	.894	Beam Theory
		.873	.620	.710	.749	Catalysis
*		.835	.310	.902	.835	Self-Consistent Field Theory
		.799	.575	.846	.749	Fluctuations and Brownian Movement
*		.132	.179	.645	.551	High Energy Nuclear Chemistry
		.628	.499	.735	.728	Dipole Moment and Polarizability
		.911	.900	.789	.815	Drugs and Poisons
*		.872	.937	.954	.927	Biosynthesis
		.916	.954	.516	.343	Structural Polysaccharides
	o	.920	.953	.831	.660	Simple Lipids
*		.955	.920	.224	.905	Enzymes
		16.7	15.8	17.6	19.2	Common Variance
		22.0	22.0	22.0	24.0	Total Variance

Table 12

Stability of "Non-Empty" Fields

Analysis											Field
A	B	C	D	E	F	G	K	L	M	N	
.806	.818	.837	.877	.859	.810	.830	.808	.773	.828	.866	1. Electric Machinery
.802	.708	.723	.522	.630	.714	.675	.844	.763	.778	.544	3. Instrumentation
.522	.552	.456	.426	.642	.508	.569	.418	.442	.548	.630	5. Field Theory
.800	.806	.808	.852	.821	.807	.807	.840	.953	.833	.853	7. Power Gen. & Distr.
.881	.895	.921	.885	.888	.896	.898	.925	.951	.923	.912	10. Aircraft Structure
.848	.871	.800	.863	.862	.890	.884	.836	.734	.895	.893	11. Aerodynamics
.450	.712	.730	.543	.746	.710	.722	.879	.470	.808	.836	16. SCF Theory
.541	.533	.335	.652	.518	.627	.599	-.047	-.099	.392	.545	18. HENC
.921	.911	.876	.923	.912	.904	.914	.886	.854	.852	.923	21. Biosynthesis
.907	.913	.867	.887	.904	.899	.903	.951	.928	.437	.909	24. Enzymes

Table 13

Stability of "Empty" Fields

Analysis											Field
A	B	C	D	E	F	G	K	L	M	N	
.735	.638	.677	.505	.577	.677	.601	.548	.527	.750	.431	4. Radar
.830	.749	.769	.754	.703	.782	.754	.672	.832	.789	.777	6. Audio Engineering
.689	.433	.612	.539	.583	.600	.610	--	--	.400	.635	8. Solid State Engr.
.742	.784	.598	.800	.725	.764	.740	.521	.278	.727	.764	12. Aircraft Design
.844	.821	.787	.829	.817	.812	.882	.252	.661	.873	.844	13. Air Properties
.873	.878	.832	.872	.864	.884	.879	.932	.862	.921	.915	14. Beam Theory
.623	.645	.264	.688	.527	.584	.557	.899	.739	.631	.548	15. Catalysis
.415	.606	.564	.420	.487	.587	.536	--	--	.662	.556	17. Fluctuations etc.
.672	.720	.717	.810	.804	.790	.796	--	--	.705	.798	19. Dipole Moment
.782	.813	.799	.815	.760	.841	.815	.855	.842	.676	.827	20. Drugs & Poisons
.502	.500	.495	.578	.458	.572	.515	.775	.821	.680	.540	22. Structural Polysaccharides
.612	.637	.718	.709	.649	.708	.694	.680	.434	.457	.734	23. Simple Lipids

Likelihood method was used in the latter. In Table 10, the fields associated with each factor are listed in the order of decreasing loadings in Analysis N, which is considered to represent the most valid set of results. In Table 10 and Table 11, the fields from which the terms for SE-III, -IV and -V were drawn are identified by asterisks; those fields which were represented by fewer than three informants are identified by zeroes.

The major discrepancies between the four analyses of Table 10 and 11 are the following:

(1) Fields which appear on Factor I for the M and N analyses split into two groups of fields on both the L and K analyses.

(2) Fields which appear on Factor III for the L, M, and N analyses split into two groups in the K analysis.

(3) Minor factors VII, VIII, and IX are common to the M and N analysis but do not appear in the L and K analyses.

(4) Minor factor XI appears in the L and K analyses but not in the M and N analyses.

(5) Minor factor X appears in the K, L, and M analysis but not on the N analysis.

(6) Fields 5 and 18 (Field Theory and Henc) show a marked drop in communality for the K and L analyses as contrasted with the M and N analyses.

(7) Field 24 (Enzymes) shows a marked drop in communality in the M analysis only.

(8) Field 22 (Structural Polysaccharides) shows a marked increase in communality for the K, L, and M analyses as contrasted with the N analysis.

On the other hand, the following facts also stand out:

(1) The six major factors found in SE-I are found in all of the present analyses.

(2) If we restrict our attention to the ten fields, identified by the asterisks, from which all the terms in the K, L, and M analyses were drawn, we find that

(a) The major projections of these fields (identified in Table 10 by double asterisks) show a degree of correspondence which is quite comparable to the correspondence found for these fields in SE-I and indeed, may well exceed the latter. The sole exception is High Energy Nuclear Chemistry (see 6, above)

(b) These fields are almost exclusively the most important determiners of the six major factors.

(3) The L analysis does not differ from the M and N analyses more than the K analysis does.

(4) The M analysis is in general, intermediate between the N analysis on the one hand and the K and L analyses on the other.

(5) The K and L analyses differ more than any two of the SE-I analyses.

These observations lead to the following conclusions:

(1) The change of corpus does not appear to have been a significant source of instability, since the L analysis does not differ from the N analysis more than the K analysis does.

(2) Although an effect resulting from the change of informants cannot be ruled out straight off, and, in fact, does offer one explanation for

Field 18, Henc, the difference in stability for the fields shown in Table 13 as compared with those in Table 12, makes it very unlikely that the change of informants as such was a major source of instability.

(3) In contrast, all of the results are consistent with the conclusion that the only significant source of instability was the combination of artifacts, already mentioned, involving the sampling from fields and the appropriate number of informants per field. The fact that the ten fields which were represented by terms (identified in Table 12 as "non-empty" fields) in the K, L, and M analyses gave results which are virtually indistinguishable from the results of SE-I whereas the remaining fields, taken as a whole, showed major discrepancies is one of the strongest pieces of supporting evidence. So, too, is the fact that the results of analysis M, where the sampling artifact is the only source of discrepancy relative to analysis N, are intermediate between the results of analysis N and analyses K and L. This comparison demonstrates that the kind of discrepancy observed between analysis N and analyses K and L can be produced by means of heavily biased sampling of terms from fields. Even the kind of discrepancy shown by High Energy Nuclear Chemistry, i.e. the radical change in communality (Table 11), in the K and L analyses is duplicated by Enzymes and Structural Polysaccharides in the M analysis.

2.3.6 Summary of the Stability Experiments

The Stability Study was designed to provide evidence in regard to the stability of factor results when the following parameters of the C-Space paradigm were altered, either singly or in various combinations (See Table 3).

(a) The size of the sample of terms rated by informants

(b) The specific set of informants used

(c) The number of informants used

(d) The specific set of terms used

(e) The specific set of references drawn from the literature

(f) The method of factoring used in the data analysis

The additional parameter,

(g) The uniformity of sampling of field corpora in the selection of terms was investigated as a result of the extraneous alteration of the experimental situation in this respect in the course of evaluating the other parametric changes.

It was found that the number of informants per field made a very substantial difference in the factor results. The use of three informants per field was found to give a stable and conceptually coherent configuration of fields in the factor space; in contrast, the use of individual informants produced a confused and largely unusable (for indexing or measuring) configuration in the factor space. The use of a three-informant minimum in the C-Space Study was considered to be vindicated by this finding.

There was an ambiguous indication that changing from one set of informants to another might occasionally make a substantial difference in the location of a field in the factor space. However, this evidence was also explicable in terms of the uniformity of sampling, and there was no indication that a change in informants was a substantial influence in producing change.

There was convincing evidence that failure to sample terms properly from the set of fields analysed produces some very marked changes in the configuration of the field vectors in the factor space. Since the kind of sampling bias which produced these effects could not arise in the normal C-Space procedure, but was an artifact of specific conditions in two of the stability experiments, the main value of this finding was to prevent the attribution of the changes that did occur to the effect of the other parameters under investigation.

The major finding of the Stability Study was that under the conditions which obtained in the C-Space Study, including a 3-informant minimum and uniform sampling of terms from fields, the factor results are virtually invariant with respect to kinds of change which in most comparable experimental contexts would have very substantial effects on empirical outcomes. Under the simultaneous alteration of all six of the parameters a-f (Analysis L vs any of analyses A-F), there was still no convincing evidence of any effect other than the effect of parameter g, sampling bias (note that here the change in parameter c differed substantially in level and degree from SE II, where number of informants was shown to be an important variable).

2.4 The Vocabulary Study

More is required for a functional IS and R system than a descriptive indexing schema. Primarily, what is required in addition, is an indexing process and a retrieval process. That is, a set of specific, realizable operations are needed which, when applied to a document, result in the document's being indexed or, when applied to a retrieval request, result in a set of documents being retrieved in response to the request. Provisions for a retrieval process were described in connection with the C-Space rationale, and empirical findings relevant to this aspect of the problem are presented in Section 2.5. The present study is concerned with the indexing process.

Certain constraints on possible indexing processes are imposed by the acceptance of the condition that the process be fully automatic. For example, the C-Space procedures would serve the purpose of indexing documents--this would require merely that the documents be rated in the same way that the C-Space technical terms and sentences were. This process would produce results which were maximally valid from the point of view of the rationale presented in Section 2.1. Such a procedure, however, would be quite unwieldy--though perhaps not more so than some methods now in use--and certainly would not meet the criterion of complete automation. Thus, the indexing problem may be rephrased as the problem of estimating, via some process which is fully automatic, the C-Space classification of documents that would have been obtained using expert informants in the C-Space paradigm.

A suggestion toward this end may be drawn from the statements that (2.1-7) the degree of relevance of a molar discursive unit to a given field is a function of the degree of relevance of its constituent statements, and (2.1-9) the degree of relevance of a technical expression to a given field is a function of the degree of relevance to that field of the statements in which the term in question does or can appear. The suggestion is best expressed in the form of a maxim: "The degree of relevance of a statement or larger discursive unit with respect to a given field can be expressed as a function of the degree of relevance of its constituent technical terms to that field."

Proceeding in accordance with this maxim, we may then try to identify specific functions which will permit technical expressions to be used effectively as estimators for classifying the discursive units in which they appear. In the Vocabulary Study, one such functions is investigated, with the effectiveness of the estimation being determined by reference to the criterion of direct psychometric C-Space indexing.

A further issue, not unrelated to the question "What function?" is raised by the questions, "Which terms?" and "How many terms." The issue is a critical one because the number of technical terms in current use in any field of knowledge is likely to be quite large (vide 2300 terms in just the 36-paragraph corpora for 20 C-Space fields); moreover the pool of current terms is being augmented and reduced at an appreciable rate, though fortunately not ordinarily at a fast enough rate to make the pool substantially unstable. Thus, any estimating procedure in which it was essential to make use of all or nearly all the technical expressions in a document in order to index the document effectively would be certain to fail in a substantial proportion of cases--probably most cases--and it would tend to be both extremely costly and extremely limited in its range of practical applicability. Thus an appropriate procedure for evaluating an estimation function would be by means of a performance curve showing indexing accuracy in relation to the number or proportion of constituent technical expressions required to achieve that degree of accuracy. This was the approach taken in the present study.

2.4.1 Procedures

From the six-paragraph corpus for each of the ten fields listed in Table 8, one paragraph was selected at random. For each paragraph, all or most of the technical expressions identified in the paragraph were rated in regard to their degree of relevance to the 24 fields listed in Table 4. The informants for the present study were the same as the informants for Stability Experiment I. The data was collected from these informants along with the data for SE-I.

For this study, the results of SE-V Analysis N were used as a C-Space. Given the ratings of paragraphs and constituents with respect to the 24 fields, C-Space coordinates were computed, so that each paragraph and each constituent term was assigned to a specific location in this miniature C-Space comprising six common factors and eight "unique" factors. The following formula was used for computing coordinates:

$$X_{Ki} = \frac{\sum A_{ij}^{3} R_{Kj}}{\sum A_{ij}^{3}} A_{if} + 0.50$$

where X_{ki} = the computed coordinate value of unit K (a term or a paragraph) on the ith reference axis

A_{ij} = the factor loading of the jth field on the ith reference axis, with j ranging over those fields used as estimators for i.

R_{kj} = the rated degree of relevance of unit K to field j

f = the one field having the highest loading on the ith reference axis

The use of this formula provides a simple weighted-average estimation of coordinate values

(a) with substantially greater weight being given to fields having higher as against lower factor loadings on the reference axis in question

(b) in a C-Space having essentially the same metric as the rating scales, i.e. a range from 0.0 to 8.0

(c) except that the upper bound for coordinate values is not 8.0 but rather that proportion of 8.0 given by A_if

(d) the constant 0.5 being added in order to avoid problems of computer underflow in the application of the estimation function.

A variety of estimation functions were defined. The major analysis was carried out with Classification Formula Three:

$$X_{ip} = (A_{ip} + B_{ip}) / 2.0$$

$$A_{ip} = \sum_{i=1}^{N} A_{iKp} / N$$

$$B_{ip} = 8A_{if} \, (A_{i1p} A_{i2p} \cdots A_{iNp}) / \sum_{i=1}^{r} (A_{i1p} A_{i2p} \cdots A_{iNp})$$

where X_{ip} is the computed coordinate value of paragraph p or reference axis i

N is the number of constituents used as estimators

K_p is the Kth constituent of paragraph p

A_{ikp} is the "known" coordinate value of K_p on i

A_{if} is defined above

r is the number of reference axes in the C-Space

More discursively, A_{ip} is the average of the i-coordinate values of the constituents and B_{ip} is the product of the i-coordinate values of the constituents, normalized, first, with respect to the sum of these products over the 14 reference axes, and second, with respect to the metric of the C-Space. The formula was selected on the basis of the functional properties of A_{ip} and B_{ip}. The first, being a simple average, tends to preserve the effects of single occurrences of substantial projections of constituents on a given axis. The second, being a product ratio, is a measure of preponderance of substantialprojections on one axis rather than another; it reflects consistency rather than single occurrences and, if used alone, tends toward an all-or-none pattern of a maximum value on one axis and essentially zero values on the remaining axes.

In general, the differences among the various Classification Formula which were defined had to do with (a) the speed with which the preponderance effect became dominant with increasing number of constituents, (b) the rapidity and boundary values for damping the preponderance effect, and (c) ways of identifying secondary nodes of relevance in conjunction with an overall preponderance effect. It is clear that the evaluation of a variety of Classification Formulae is a significant area for further investigation.

In the sequential procedure in which first one, then 2, . . . then N constituents were used for estimating paragraph locations, each set of estimators represented a separate random selection from the total available. This procedure has the limitation that as N approaches the total number of constituents, the successive samples become less and less independent.

2.4.2 Results

The results of this analysis are shown in Table 14. There are two major findings. The first is that the Classification Formula used in the study succeeds quite well in matching the criterion locations of the paragraphs. The degree of discrepancy reflected in the distance measures shown Table 14 is to be interpreted in the light of a 14-dimensional C-Space in which the greatest distance from any allowable indexing location to the origin is 22.56 and in which essentially all of the data falls in the positive manifold. The second is that although there is some tendency for indexing accuracy to increase as more and more constituent terms are used, there is essentially no increase in accuracy past the range of 3-6 terms.

Thus, if the level of accuracy attained by this estimation function is adequate for use in an operational setting, the present results provide a strong basis for expecting that the number of previously indexed technical expressions required for processing documents in a given field will be smaller, by one to two orders of magnitude, than the total number of technical expressions having current use in the field.

The present results may be interpreted as reflecting the same basic phenomenon as the Stability Study, i.e. that the use of the concept of relevance in relation to subject matter shares with the exercise of most human capacities the characteristic that its operation under optimal conditions is surprisingly crude in comparison with the effectiveness with which it operates under minimally favorable conditions. The results of both studies are more readily assimilable to the threshold-sensitivity paradigm of basic perceptual processes than to the decay-function paradigm of acquired abilities, operating

Table 14A. Estimation of Paragraph Locations: Each occurrence counted

No. Terms Used	Average Discrepancy	Range of Discrepance	Based on N Paragraphs
1	3.39	1.36 - 6.54	8
2	2.76	1.90 - 4.31	8
3	2.76	1.94 - 3.72	8
4	3.12	1.82 - 5.84	8
5	2.90	1.80 - 5.01	8
6	3.13	1.80 - 6.30	8
7	2.72	1.56 - 3.69	8
8	2.36	1.19 - 3.36	6
9	2.51	1.44 - 3.99	5
10	2.48	1.42 - 3.00	4
11	2.81	1.52 - 4.42	4
12	2.56	1.38 - 3.78	3
13	2.49	1.35 - 3.67	3
14	2.95	2.37 - 3.51	2

Table 14B. Estimation of Paragraph Locations: Each term counted once

No. Terms Used	Average Discrepancy	Range of Discrepancy	Based on N Paragraphs
1	3.47	1.36 - 6.54	8
2	3.13	1.90 - 4.52	8
3	2.96	1.61 - 4.43	8
4	3.13	1.57 - 5.11	8
5	2.73	1.63 - 4.28	8
6	3.05	1.62 - 5.78	8
7	2.72	1.35 - 3.34	8
8	2.48	1.18 - 3.29	6
9	2.63	1.46 - 3.49	5
10	2.52	1.44 - 2.90	4
11	2.59	1.38 - 3.72	4
12	2.44	1.36 - 2.58	3
13	2.43	1.32 - 3.31	3
14	2.95	2.67 - 3.24	2

with environmental support. Such comparisons are, of course, merely suggestive at the present time.

Aside from the possibility that other more effective estimation functions analogous to CF 3 may be found upon further investigation, C-Space technology affords indexing resources of a different genre which bear obvious mention. For example, the operation of a Classification Formula is independent of the linguistic unit to which it applies; in principle, this fact provides a direct solution to the problem of indexing documents which are heterogeneous in content, since a variety of techniques are available for identifying substantial shifts in content emphasis in the course of a linear progression from the beginning to the end of a given document, and documentary subunits delineated on the basis of the shifts could be indexed separately in addition to the summary indexing of the document as a whole. Again, since the indexing of a document of any substantial length may be regarded as both a succession and a cumulation of sub-unit indexings, a document may be described in terms of its trace and its volume in C-Space and, for example, the ratio of the total volume to the cylindrical volume around the trace (i.e. a sort of multidimensional regression line) might well serve to distinguish a general discussion of a broad content area as against the successive discussion of several specialized areas when the summary classification of both documents is the same.

2.5 The Relevance Ranking Study

The procedure adopted for sequential retrieval from a C-Space is based on the statement (Section 2.2-h) that when one of two discursive units indexed in a C-Space has the pragmatic status of a retrieval request and the second is a sentence or larger unit (i.e. a candidate for retrieval), the C-Space distance from the second unit to the first is a monotonic function of of the degree of relevance of that unit to the retrieval request. Considering that a C-Space is constructed on the basis of the degree of relevance of discursive units to fields of knowledge, this assertion may appear paradoxical, as if, for example, one were to say that if two keys fit the same set of locks, then one of the keys fits the other. The puzzle, however, arises from the failure to take account of the pragmatic status of a retrieval request. Although a retrieval request can be associated with a point in the C-Space on the basis of the index values of its constituent technical expressions (it may itself consist merely of a list of one or more such expressions), it is not a piece of information which is a candidate for retrieval, so that it does not have the same kind of relevance to a set of fields of knowledge that the literature of those fields has. The kind of relevance it does have is already directly expressed and distinguished from other kinds of relevance as soon as it is characterized as a "retrieval request."

The pragmatic function of a retrieval request is to delimit a range of subject matter (of the "criterial description" referred to in Section 1.1-a). This is likewise one of the functions of "field of knowledge" concepts. Thus, a retrieval request may be assimilated to the concept of a "possible field of knowledge" (vide Section 2.2-j). On this view, the coordinates computed for a retrieval request are not to be interpreted as defining a point-location for the request. Instead, these coordinates determine the projection of the corresponding "possible field" into the C-Space. Here, the same considerations apply as in connection with the problem of unique content (Section 2.2-d) of C-Space fields. To the extent that the point-location associated with the request is close to the outer limit of permissible C-Space locations, i.e. is far out from the origin, the content of the "possible field" is fairly exhaustively contained within the scope of the C-Space, and consequently, sequential retrieval determined by distance from the request would approach maximal correspondence to the order of degree of relevance to the request. However, to the extent that the location associated with the retrieval request is close to the C-Space origin, the "possible field would have a considerable proportion of its subject content not represented in the C-Space; consequently all of the material (documents) in the C-Space would tend to have a relatively low degree of relevance to the "possible field" and sequential retrieval according to the distance of documents from the C-Space location of the request would be relatively ineffective .

This state of affairs may be viewed in several lights. For example, it indicates a kind of limitation on the effectiveness of the C-Space technology. Or, the notion that the C-Space, like any technology, has limitations, may be taken for granted and attention focussed on the significant advantage of the fact that the C-Space technology itself provides a discriminating measure of its own limitations with respect to any given retrieval request. Or again, it may be seen as an argument for the development of a "complete" C-Space, i.e. one which is not limited in the scope of its content and therefore lacks the limitation in question. This latter comes close to merely repeating that the appropriate conceptual context for dealing with subject matter is a pragmatic context, i.e. the widest possible context in which the concept of "subject matter" appears in its own right and not merely as a factual instance of other concepts.

Nevertheless, the articulation of pragmatic concepts is sufficiently unfamiliar, and the delineation of this particular aspect is sufficiently complex, so that one would like to see some relevant evidence on the matter. The Relevance Ranking Study was designed to provide such evidence.

2.5.1 Procedures

Two fields from each of the four major content areas were selected for study. These fields, listed in Table 15, were selected from the ten fields (Table 8) for which two sets of references, i.e. a total of 72 paragraphs, were available. For each major content area, one field was designated as the prime field and the other field was designated as the secondary field. As shown in Table 16, three paragraphs were selected at random from the corpus of each prime field and one from each secondary field. For each paragraph, a subject matter title, assumed to be at least roughly descriptive of the content of the paragraph, was constructed. In all cases, the paragraph title was the title of the reference or subheading under which the paragraph appeared in the original text, or else it was a close paraphrase thereof.

Two sets of informants were involved. Four of the paragraphs had been rated directly by the Vocabulary Study informants, and so these ratings were incorporated directly as part of the Relevance Ranking data. Ratings

Table 15

Relevance Ranking Study Fields

Field	
FT	Field Theory
EM	Electric Machinery
Ad	Aerodynamics
AS	Aircraft Structures
HENC	High Energy Nuclear Chemistry
SCF	Self-Consistent Field Theory
B	Biosynthesis
SL	Simple Lipids

Table 16

Paragraph Titles

Field	Paragraph	Title
FT	C	Vector Analysis
FT	K	Types of Fields
FT	A	Harmonic Functions
EM	R	Construction of a Power Generating Plant
Ad	J	Parameters of Aerodynamic Forces and Moments
Ad	H	Lift Analysis
AS	V	Wing Theory
Ad	S	Contraction Properties
HENC	B	Atomic and tonic recoil from the ( ) reaction
HENC	X	Nonreacting Collisions of Energetic Recoil Atoms
SCF	Q	Radiolysis of Propane - d_8
HENC	U	Gamma Ray Emission
B	W	The Relation of Nucleic Acids to Proteins
SL	G	The Synthesis of Fat
B	M	How a Protein is Made
B	D	The Synthesis of Sugars and Sugarlike Compounds

on the remaining twelve paragraphs and on the sixteen titles were obtained from the group of informants who participated in Stability Experiments III and IV.

As was done in the Vocabulary Study, a 14-dimensional C-Space based on the 24-field, 288-term Maximum Likelihood analysis (SE-V Analysis N) was used. Likewise the formula used in the Vocabulary Study for computing coordinates on the basis of ratings with respect to the 22 fields was used in the present study.

Eight groups of six paragraphs were assembled, as shown in Table 17. The paragraphs were printed on separate cards and the set of cards was shuffled prior to being presented to an informant. The instructions to the informants are shown in Appendix B. Essentially, they were given one of the paragraph titles and asked to consider this a "topic" about which they might have gone to find information, and then to rank the six paragraphs in their order of relevance to the topic. Table 17 shows that each set of topic plus paragraphs had the following characteristics.

(a) The topic was identical with the title of one of the paragraphs in the set, and that paragraph was selected from one of the prime fields.

(b) Of the six paragraphs in the set, three were from the major content area of the paragraph the title of which was the "topic" associated with the set. The three remaining paragraphs included one paragraph from each of the other three major content areas. (This procedure was designed to ensure a substantial range of relevance with respect to the topic, and also to ensure broad sampling from the four major content areas.)

It was also the case for each set that the relevance rankings were made only by informants whose field of special competence was the field corresponding to the paragraph from which the "Topic" was taken. Thus, the rankings were performed by informants from Field Theory (2 informants), Aerodynamics (4), High Energy Nuclear Chemistry (3), and Biosynthesis (6).

2.5.2 Results

The major results of the study are shown in Table 17. The average rank of each paragraph in relation to the topic is shown under the heading "Judged Relevance" and the C-Space distance between the location associated with the Topic and the locations of the paragraphs is shown in the adjacent column. It is evident that in general there is a very high degree of correspondence between, on the one hand, the degree of relevance of the paragraphs in the C-Space to the "retrieval request" and, on the other hand, the respective distances from the C-Space location of the paragraphs to the C-Space location of the "retrieval request". Figure 1 indicates that the relation of relevance to distance is not merely monotonic, but is highly linear as well.

The major exception to the foregoing is the set associated with the topic "Contraction Properties. This is the set connected by lines in Figure 1. It can be seen that the exclusion of this set leaves a scatter plot which represents an even greater degree of relationship than the original. This set also provides the single exception to the finding that with respect to a given Topic, the three paragraphs judged to be most relevant are the three from the major content area appropriate to the Topic. If we examine the locations of the Topics in regard to their distance from the C-Space origin (Table 18), we see that the topic "Contraction Properties", in relation to which there is the least correspondence between distance and relevance and the six paragraphs are about equally relevant, with none very relevant, is located much closer to the C-Space origin than any of the other Topics and is essentially unrelated to any of the C-Space reference axes. This is the result that would be predicted on the basis of the earlier discussion of the "unique content" limitation.

One result, which is evident in Table 18 and would not have been predicted straight off is an apparent threshold effect. The correlation of distance with relevance is not a linear function of the distance of the Topic from the origin. Instead, there appears to be a critical value below which C-Space distance does not parallel relevance ranking and above which it does. It is not clear whether the critical value relates more to the distance of the topic from the origin or to the exteh of its greatest projection on any axis. If the threshold effect were a consistent finding this fact would tend to maximize the likelihood of automatically identifying retrieval requests which could not be adequately responded to by the system.

Table 18 also shows that, with the exception of "Contraction Properties", the correlation between C-Space distance and the average judged relevance of paragraph to topic is almost uniformly greater than the average correlation between the judgments of relevance made by the individual informants.

An evaluation of the results of the Relevance Ranking Study should take into consideration the degree to which inter-individual contingencies entered into the outcome. For example, the C-Space itself was based on one set of informants and almost all of the locations of linguistic units were based on a different set of informants. Similarly, although the relevance of all six paragraphs relative to a given Topic were made by single informants, the computed distances from the topic to the six paragraphs were based on ratings from approximately 26 other informants.

On the whole, it appears that the present results lend considerable support to the proposition that a well-constructed C-Space does in fact have the general relevance properties which were attributed to it as a result of interpreting some pragmatic concepts within the limits of the methodology of factor analysis and factor measurement. The present results support the notion that these relevance relationships mapped by a C-Space are relatively stable social realities rather than idiosyncratic individual opinions.

Figure 2 shows the results obtained when the relevance rankings are compared with distances based on automatic indexing, using the Classification Formula on technical expressions occurring in the paragraphs which were ranked. The number of terms used for indexing a given paragraph ranged from four to six. As might be expected, these results are somewhat less impressive than the results for psychometric indexing shown in Figure 1, but there is not a great deal of difference in the two sets of results. The outcome shown in Figure 2 bears out the indications found in the Vocabulary Study that effective indexing could be obtained with a surprisingly small number of terms. However, the primary significance of the present findings is not that a specific level of performance has been achieved at first trial, but rather, that they close the last major gap in the indexing--storage--retrieval sequence provided by a C-Space system. In conjunction with previous findings, the present results indicate that there are no seriously weak links in the basic process (this does not eliminate the possibility of serious problems arising from boundary conditions). Considering the diversity of concepts, procedures, and empirical influences which collectively make up the C-Space package, this provides a good deal of motivation and background confidence for exploring and developing the potential of this approach to LDP problems.

Table 17

Paragraph Sets

Topic: (C) Vector Analysis

Paragraph	Title	Judged Relevance	C-Space Distance
K	Types of Fields	1.00	1.89
A	Harmonic Functions	2.00	2.65
C	Vector Analysis	3.00	2.35
H	Lift Analysis	4.00	5.02
Q	Radiolysis of Propane - d_8	5.25	7.72
M	How a Protein is Made	5.75	8.74

Topic: (K) Types of Fields

Paragraph	Title	Judged Relevance	C-Space Distance
K	Types of Fields	1.50	1.69
R	Construction of a Power - Generating Plant	2.00	4.91
C	Vector Analysis	2.50	2.35
B	Parameters of Aerodynamic Forces and Moments	4.50	6.25
J	Atomic and Tonic Recoil	4.50	7.41
W	The Relation of Nucleic Acids to Proteins	6.00	8.26

Topic: (B) Atomic and Tonic Recoil

Paragraph	Title	Judged Relevance	C-Space Distance
B	Atomic and Ionic Recoil	1.00	.65
X	Nonreacting Collisions of Energetic Recoil Atoms	2.00	1.41
Q	Radiolysis of Propane - d_8	3.60	1.96
V	Wing Theory	4.30	6.46
K	Types of Fields	4.60	7.32
G	The Synthesis of Fat	4.60	8.61

Topic: (X) Nonreacting Collisions of Energetic Recoil Atoms

Paragraph	Title	Judged Relevance	C-Space Distance
X	Nonreacting Collisions of Energetic Recoil Atoms	1.30	1.37
Q	Radiolysis of Propane - d_8	2.00	1.92
U	Gamma Ray Emission	2.60	3.76
S	Contraction Properties	4.60	5.60
D	Synthesis of Sugars and Sugar-like Compounds	5.00	9.18
R	Construction of a Power-Generating Plant	5.30	8.44

Table 17 (continued)

Topic: (W) The Relation of Nucleic Acids to Proteins

Paragraph	Title	Judged Relevance	C-Space Distance
W	The Relation of Nucleic Acids to Proteins	1.4	1.07
M	How a Protein is Made	1.6	.79
G	The Synthesis of Fat	3.0	2.50
V	Nonreacting Collisions of Energetic Recoil Atoms	4.6	6.14
X	Wing Theory	4.6	7.01
K	Type of Fields	5.6	7.71

Topic: (G) The Synthesis of Fat

Paragraph	Title	Judged Relevance	C-Space Distance
G	The Synthesis of Fat	1.16	3.48
D	Synthesis of Sugars and Sugar-like Compounds	1.84	2.07
M	How a Protein is Made	3.00	3.16
V	Gamma Ray Emission	4.50	9.96
R	Construction of a Power-Generating Plant	5.16	10.78
S	Contraction Properties	5.30	8.57

Topic: (J) Parameters of Aerodynamic Moments and Forces

Paragraph	Title	Judged Relevance	C-Space Distance
J	Parameters of Aerodynamic Forces and Moments	1.75	1.51
H	Lift Analysis	1.75	2.53
V	Wing Theory	2.50	2.67
C	Vector Analysis	4.00	4.72
Q	Radiolysis of Propane - d_8	5.25	6.22
M	How a Protein is Made	5.75	7.60

Topic: (S) Contraction Properties

Paragraph	Title	Judged Relevance	C-Space Distance
S	Contraction Properties	1.75	3.42
V	Wing Theory	2.50	4.05
A	Harmonic Functions	2.75	6.07
H	Lift Analysis	3.00	4.44
B	Atomic and Ionic Recoil from the (η, γ) reactions	5.00	4.54
W	The Relation of Nucleic Acids to Proteins	6.00	4.56

Table 18

"Uniqueness" of Request

Topic	Distance from Origin	Correlation with Criterion	Highest Factor Loading	Average Correlation of Individuals
Vector Analysis	8.28	.949	4.44	.986
The Synthesis of Fat	8.20	.955	5.56	.752
Types of Fields	7.54	.900	4.44	.771
Relation of Nucleic Acids to Proteins	6.03	.913	4.88	.695
Aerodynamic Forces and Moments	6.00	.984	4.20	.848
Atomic and Ionic Recoil	5.23	.981	4.69	.881
Nonreacting Collisions	5.22	.896	4.52	.870
Contraction Properties	2.64	.196	1.16	.686

Figure 1

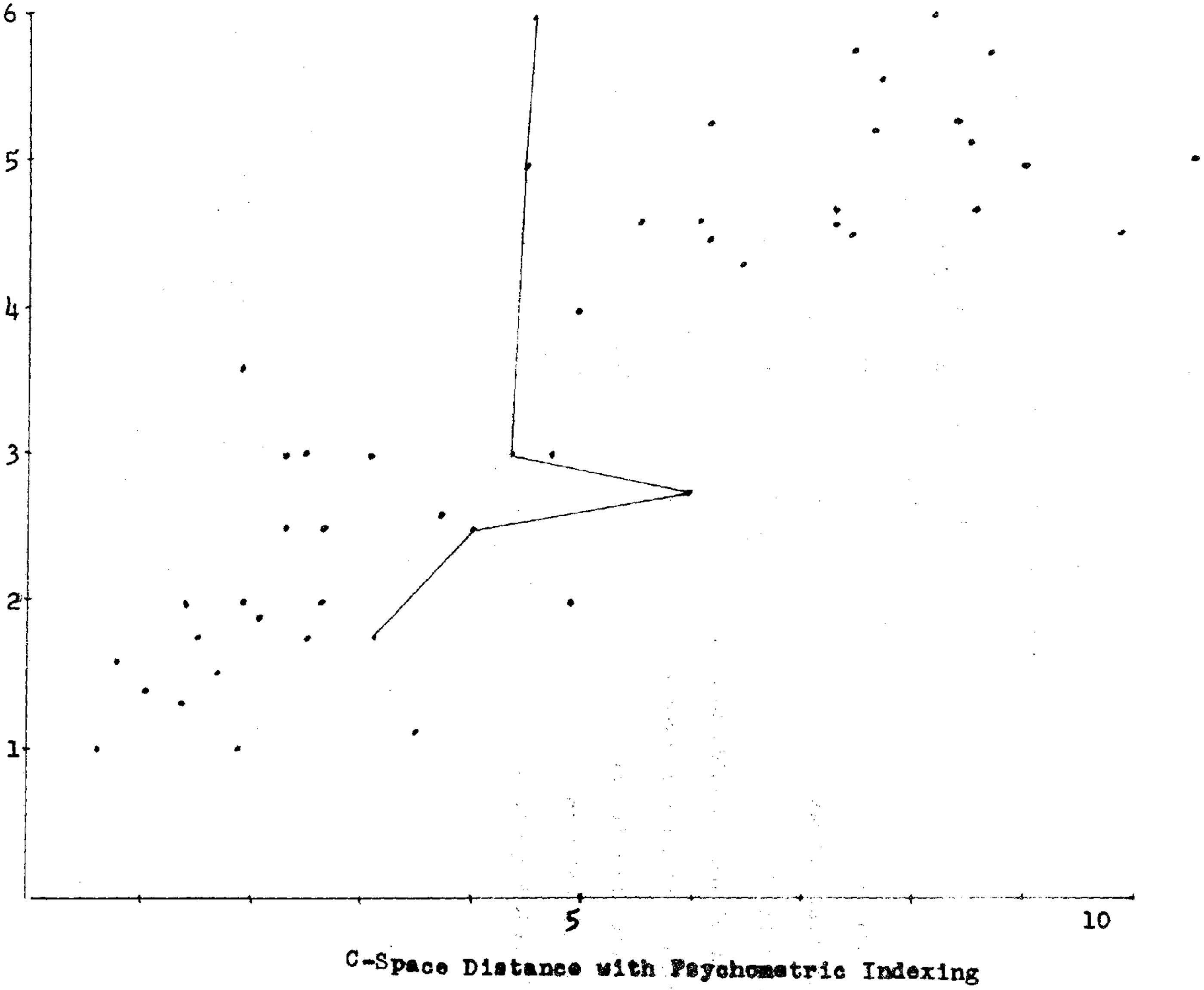
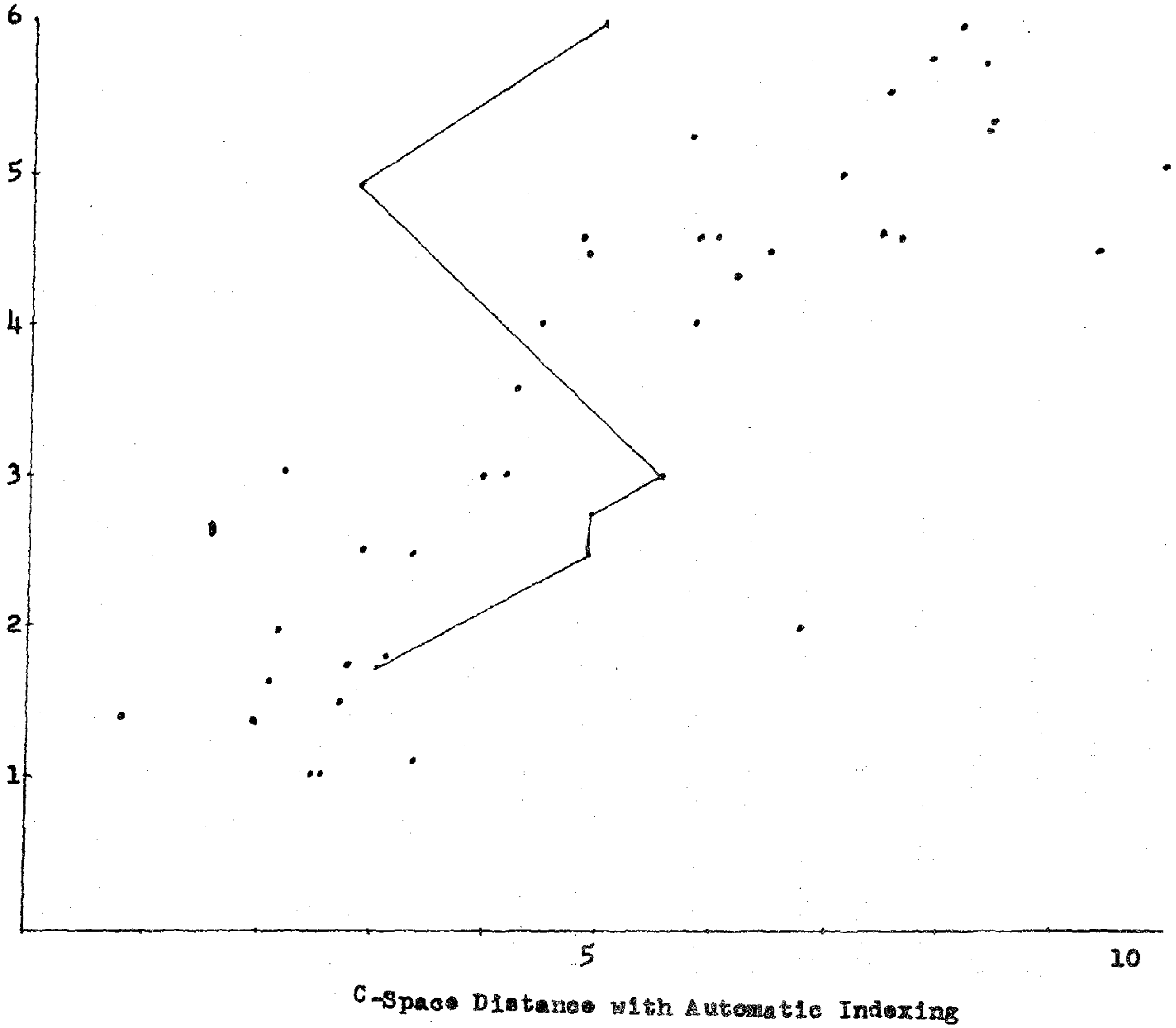

Relevance Rank
1
2
3
4
5
6
5
10
C-Space Distance with Psychometric Indexing

Figure 2


Relevance Rank
6
5
4
3
2
1
5
10
C-Space Distance with Automatic Indexing

2.6 The Grammatical Study

One major focus in the development of the C-Space technology would be the improvement of the Classification Formula for automatic indexing. It is also the case that it is a priori plausible that the syntactic function of a given technical expression in a given occurrence should have a bearing on the information potential of the term so used. Certain general, informal observations contribute to this impression, for example the observation that "the important words come first in a sentence." In conjunction with the further observation that in English sentences, subject terms have a strong tendency to occur before object terms, this notion leads directly to the hypothesis that the accuracy of automatic indexing will be improved if technical expressions which occur as subjects are given greater weight than expressions which are functioning as objects. In the present study the clause subject or clause object status of technical expressions was used as a basis for differential weighting in an attempt to provide increased accuracy of automatic indexing.

2.6.1 Procedures

Two sets of twenty sentences each were used as a corpus. In each sentence the constituent technical terms were identified and their subject or object status was established. Using the same instructional setting and informants as Stability Experiment I, ratings were obtained with respect to the twenty-four fields of SE-I for each sentence and either all or most of the constituent expressions for each sentence. C-Space coordinates were computed for each sentence and technical expression by means of the procedures described in the report of the Vocabulary Study. The C-Space location of each sentence was then estimated on the basis of the location and grammatical status of its constituent technical terms. For this purpose a modification of the Classification Formula used in the Vocabulary Study was used. The modification consisted of substituting Tan $(K A_{ij})$ in place of the simple coordinates A_{ij} for subject expressions and Tan $(H A_{ij})$ in place of the simple coordinates for object expressions. The choice of the tangent function represented an attempt to manipulate the "preponderance effect" discussed earlier. Simple multiplication by a constant was ruled out by the nature of the Classification Formula, which automatically eliminated the effect of such multiplications. The use of a power formula, it appeared, would produce too much of a preponderance effect to be studied in a sensitive way. The tangent function appeared to provide an easily controlled effect, since it is increasingly nonlinear over a wide range but neither the total departure from linearity nor the rate of change of degree of linearity is very great at any point between 0° and 90°.

Repeated calculations of sentence locations were made under systematic variation of the parameters K and H. For each parametric change, the distance between the sentence location and the estimated location was calculated. The average results are shown in Table 19a for the first twenty sentences, Set A, and in Table 19b for the second twenty sentences, Set B.

2.6.2 Results

The results in Table 19 are for values of HA_{ij} and KA_{ij} ranging from approximately 5° to approximately 55° for object expressions and

5^{o} to 45^{o} for subject expressions. The initial plan was to scan a wide range with large parametric increments and subsequently to scan a more limited range with small parametric increments, if there appeared to be a sink in the distance measure of error in some region for both sets of sentences. The continuity of the tangant function, as indicated above, was considered to provide a reasonable basis for interpolating between parametric values. However, the results on the first pass were so clearly negative that it appeared inadvisable to pursue this analysis further. Table 19a shows a uniform increase in error with increasing values of parameters H and K whereas Table 19b shows a uniform decrease in error with increasing values of H and K. It is true, however, that in both cases the subject expressions showed a greater sensitivity to the parametric changes. This result tends to support the general notion that subject expressions are somehow "more important" than object expressions.

Table 19A. Distance Error for Set A Sentences

Values of	Values of K (objects)				
H (subjects)	3	4	5	6	7
3	3.014	3.013	3.012	3.011	3.011
4	3.012	3.012	3.011	3.010	3.009
5	3.010	3.010	3.009	3.008	3.005
6	3.008	3.007	3.006	3.006	3.005

Table 19B. Distance Error for Set B Sentences

Values of	Values of K (objects)				
H (subjects)	3	4	5	6	7
3	3.526	3.527	3.529	3.531	3.534
4	3.529	3.530	3.531	3.533	3.536
5	3.532	3.533	3.535	3.537	3.539
6	3.536	3.537	3.539	3.541	3.544

3.0 Summary Discussion

A conceptual approach has been sketched and empirical illustrations presented of the major softwave components -- input, storage, output -- of a linguistic data processing system which offers, in principle, the advantages of complete automation, unlimited cross-indexing, effective sequential retrieval, subdocumentary indexing reflecting heterogeneity of subject matter, and a procedure for identifying retrieval requests which would be poorly served by the system.

The major contributions of the empirical studies described in Section 2 are the following:

(a) The construction of a C - Space having sufficient content scope to provide the indexing structure for a limited LDP system in an operational setting.

(b) The demonstration that a properly constructed C - Space is, and an improperly constructed one is not, an exceptionally stable structure.

(c) The illustration of specific computational procedures for effective automatic indexing of documents in a C - Space, using a relatively small system vocabulary.

(d) The demonstration that a C - Space does map relevance relationships and thereby promotes effective sequential retrieval.

(e) Clearcut negative findings in relation to a limited attempt to use the structural differentiation of subject and object as a means of increasing the effectiveness of automatic indexing in a C - Space, together with some indication that the grammatical status of the constituent terms does make some kind of difference in regard to their indexing power.

There can be little question that the C - Space technique shows sufficient promise to warrant a good deal of further exploration and development. Certain areas for further work are fairly clearcut:

(a) The further investigation of alternate means for automatic indexing in the C - Space. The use of structural features of the text still appears to be a promising approach in spite of the clearly negative results of the specific procedure attempted here. Or again, the indexing of a document as a function of the next smaller subunits rather than directly from the ultimate constituent terms is a very practical approach, especially in conjunction with subdocumentary indexing. And certainly, the types of estimating functions investigated already do not cover even the major possibilities along this line.

(b) Because of the uncontrolled reduction of the number of fields contributing to the currently developed C - Space, the latter will not function optimally. The extension of scope and elaboration of sampling within the present scope are desirable and straightforward procedures.

(c) A critical area for further development is the initiation of a functional C - Space LDP system. Such an arrangement, ideally, would permit the most rigorous examination of the present state of the art and at the same time would provide the most favorable conditions for identifying and analyzing significant unsolved problems. It would provide the most efficient framework for testing any additions or improvements which might be incorporated. Initially, however, certain problems not peculiar to a C - Space system, would arise simply with respect to the successful initiation of a new LDP system in a functioning organization which has already adapted to existing services.

(d) A different class of problems arises in connection with the use of factor analytic techniques on a large scale. Currently available computer programs for performing factor analyses are limited to about 125 -- 150 variables. The development of efficient programs for handling a considerably larger number would be a sizeable task in programming. In addition, there is some reason to expect that more fundamental problems relating to precision and degree of structural articulation would arise for, say, a 1000-variable analysis giving rise to a C - Space of 250 dimensions. It seems likely that under these conditions methods would have to be developed for constructing "closeups" of subregions of the space and for coordinating the closeups with the larger structure.

(e) Another class of problems, which would almost certainly be highlighted by the existence of a functioning system, has to do with the updating of the system's vocabulary and index structure in the light of changes occurring in the content domain of the C - Space.

(f) Still a different area is that of adapting the C - Space approach for use in content areas having characteristics which are significantly different from those of scientific and technical fields (e.g., the arts and humanities, law, object inventories, etc.). It is not a foregone conclusion that C - Space techniques will be equally appropriate for other content domains or that they will not be.

The very fact that so many avenues of development are open and the fact that a geometric model provides a formal context capable of accomodating indefinitely fine distinctions can easily lead to methodological excesses in the course of attempting to increase the effectiveness of C - Space indexing and retrieval. One such excess is the attempt to make finer and finer subject matter discriminations within a C - Space.

With respect to the latter, two points need to be made. The first is that (as in the case for fields of knowledge) subject matter does not have the general property of divisibility -- there are not as many subject matters as there are things that can be talked about. The second is that such refinement would be quite unnatural and is quite unnecessary so long as we are not limited to the methodological context provided by a C - Space. For human language and cognition (as contrasted with perception) it is over-whelmimgly the case that increasing differentiation of concepts is not achieved by what amounts to a purely numerical subdivision within the range of a continuous variable (or within the volume of a homogeneous N - Space) but rather, by the introduction of other descriptive contexts and the subsequent combination of the elements of these into a pragmatically complex description. ("Automobiles" refers to a distinguishable subject matter; "Blue automobiles" does not. But the former is a pragmatically simple description -- a subject matter description, whereas the latter is a complex attribute-subject matter description.)

Thus, it seems clear that the most effective way of adding to the LDP contribution of the C - Space is to devise systems for implementing other kinds of description of information. This kind of program would be particularly important considering the major likelihood that most actual (criterial) retrieval requests are only approximately expressible as simple subject matter descriptions of information. It seems likely, too, that the differences with respect to non-subject matter aspects of criterial descriptions are responsible for much of the individual differences e.g., in the ranking of documents in regard to their relevance to "the same" topic.

One of the consequences of the foregoing is that the distinction between document processing and information processing ceases to be a basic one. (This is already implicit in the introduction of automatic sub-documentary indexing.) Instead we have the basic concept of the criterial description of information and the basic distinction is that of the several kinds of such description, including subject matter descriptions. Attribute descriptions and semantic descriptions would be among the other basic kinds.

Such an approach may appear to be doomed to failure by virtue of the multiplicity of kinds of description. Certainly, the compounding of complexity has been one of the most significant and durable problems in the short history of linguistic data processing. Much, however, depends on how the "kinds" are identified. It is true that the quantity and diversity of human intellectual products is impressive. However, there is a great deal of background evidence which points to the conclusion that such products can be (and in fact, are) understood as resulting from the operation of a relatively small number of basic cognitive capacities (not "processes" or "mechanisms"). The empirical results presented in this report suggest that operating at a pragmatic conceptual level in the analysis of human activities can result in a significant degree of success in identifying and characterizing these cognitive capacities.

On the positive side, this way of reducing the distinction between document processing and information processing has the result that a wide range of LDP problems (abstracting, MT, dissemination, fact correlation, and document storage and retrieval) are seen to be systematically related in terms of the kind and range of criterial descriptions which must be implemented in their solutions.

Appendix A

C-Space Instructions

The purpose of this procedure is to obtain quantitative estimates of the degree to which a selected group of scientific and technical fields overlap in their subject matter.

This is accomplished by having people make judgments about a set of "sample items" in relation to a field of knowledge in which they are competent. The sample items include words, phrases, sentences, and paragraphs selected randomly from the literature of the fields which we are investigating.

Your basic task in rating each sample item is to decide the degree to which the content of the sample item is relevant to the field of ____________________. Another way of looking at it is that you are to decide the degree to which the content of the sample item should be regarded as a part of the subject matter of this field.

Your decision for each sample item is expressed by making a checkmark on the numerical scale which accompanies each sample item. The use of the scale is explained on the next page.

Rate each sample item independently with respect to your field. Do not try to take account of any relationship which the item may have to any other field--that will be done by people who are rating with respect to the other fields. Do not try to take account of how you have rated other sample items.

The Scales:

You will be using scales like this one:

!! ! : ! : ! : ! : !

0 1 2 3 4 5 6 7 8

In general, the more relevant the sample is to the field you are judging, the higher the number of the scale position you should mark. Use the following as a guide to the use of specific scale positions.

The sample item has no particular significance for this field; it is essentially irrelevant.

Mark !! ✓ !

0

The sample item may have some relevance to the field, but it would have to be regarded as peripheral, tangential, or incidental. Ordinarily, you wouldn't associate it with this field. Under these conditions:

If less relevant, mark ! ✓ :

1

If more relevant, mark : ✓ !

2

The sample item does have some relevance, but it is of a borderline nature. For example, the sample item might be primarily an ordinary English expression which happens to have some bearing on the content of the field; or it might fall in a "fringe" content area about which there is some question as to whether it should "really" be included in the field; or it may refer primarily to general scientific methodology rather than specifically the subject matter of this field. Under these conditions:

If less relevant, mark ! ✓ :

3

If more relevant, mark : ✓ !

4

The sample item is quite relevant to the subject matter of the field. it refers to objects, concepts, or processes, etc., which are definitely part of the subject matter of the field.

If less relevant, mark ! ✓ :

5

If more relevant, mark : ✓ !

6

The sample item is highly relevant to this field. For example, it may be a technical term representing a very refined distinction in which a great deal of the conceptual apparatus of the field is implied. Or it may be a sentence or a paragraph which mentions or implies a number of relevant concepts or which states facts or conclusions which are very significant for people in the field.

If less relevant, mark ! ✓ :

7

If more relevant, mark : ✓ !

8

Appendix B

Relevance Ranking Instructions

In each of the envelopes marked "First Envelope" and "Second Envelope" you will find a set of cards. Each card contains a paragraph of text and this paragraph is identified by the capital letter which appears above it.

I. Consider the following subject matter:

Vector Analysis

Think of it, for example, as a topic about which you might want information.

Now, take the cards from the "First Envelope" and rank order these paragraphs. Rank 1 should go to the paragraph which is most relevant to the topic or subject matter specified above. Rank 2 goes to the paragraph which is next most relevant, etc.

When you have finished, indicate your rankings here:

Rank	1	2	3	4	5	6
Para.						

(Use letter)

Try not to have any paragraphs tied for the same rank. However, if you find two which really are indistinguishable as to their relevance, indicate ties by circling the corresponding ranks. Thus, for example, here is how you would indicate that paragraphs J and Q were tied for the fourth rank:

Rank	1	2	3	4	5	6
Para.	A	X	P	J	Q	H

Appendix C

C-Space Field and Literature Identification

The following people bore the primary responsibility for the identification of fields of knowledge within the four major content areas and for the selection of the literature associated with these fields.

Electrical Engineering:	Mr. Ronald Taylor, Research Engineer
Physical Chemistry:	Mr. Peter F. Jones, Graduate Student, Department of Chemistry, UCLA
Aeronautical Engineering:	Mr. Mickey Blackledge, Graduate Student, School of Engineering, University of Colorado
	Mr. Peter Hendricks, Graduate Student School of Engineering, University of Colorado
Biochemistry:	Mr. George Dersham, Graduate Student Department of Chemistry, University of Colorado